2008

Retrotransposon-mediated instability in the human genome

Shurjo Kumar Sen
Louisiana State University and Agricultural and Mechanical College, shurjo.sen@gmail.com

RETROTRANSPOSON-MEDIATED INSTABILITY IN THE HUMAN GENOME

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Biological Sciences

by
Shurjo Kumar Sen
B.Sc. (Hons.), University of Calcutta, 2001
M.Sc., University of Calcutta, 2003
May 2008

ACKNOWLEDGEMENTS

I owe a debt of gratitude to several people for their assistance during the course of my dissertation research. My graduate advisor, Dr. Mark A. Batzer, has been no less than a father to me ever since I joined LSU, and has frequently helped me in his own special way to get over periodic attacks of laziness. Drs. Michael Hellberg, Joomyeong Kim and Stephania Cormier have been simply wonderful as members of my graduate committee. Dr. John Battista graciously spent large amounts of time answering my questions about DNA repair.

I have been lucky to have outstanding collaborators, both within and outside the Batzer laboratory whom I would like to thank for enriching my graduate school experience. For his contribution to chapters two, three and four: Dr. Kyudong Han. For their contribution to chapters two and three: Drs. Ping Liang and Richard Cordaux. For his contribution to chapter four: Charles Huang. In addition, I remain grateful to all other members of the Batzer laboratory (especially Jerilyn Walker, our many undergraduate workers, and my good buddies Jason Gray and Chad Jarreau) for being the best set of colleagues possible.

My parents, brother and extended family were a constant source of support throughout the last four years, especially my maternal grandfather who remains my only hero. Lastly, I would like to express my gratitude to my wife Soma for her love, devotion and the many delicious meals she supplied at odd hours.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

LINE-1 (Long Interspersed Element-1/L1) and *Alu* are two active retrotransposon families in the human genome that have the potential to create genomic instability either during the insertion of new elements or through ectopic recombination. However, recent *in vitro* analyses have demonstrated that these elements also repair DNA double-strand breaks, hence contributing to the maintenance of genomic integrity. As such, the comprehensive role of mobile elements in either creating or mitigating instability in primate genomes remains unclear. The recent sequencing of the chimpanzee and rhesus macaque genomes uniquely facilitates the accurate resolution of this question, as three-way computational alignment of the human genome with two other hominoid genomes allows human lineage-specific changes (i.e., those younger than 5-6 million years) to be accurately dissected out. Here, using a combined computational and experimental approach, we have attempted to provide an unbiased picture of the contribution of the *Alu* and L1 families to human genomic stability. In the first analysis described herein, we assessed levels of genomic deletion associated with L1 retrotransposition and reported 50 deletions resulting in the loss of ~18 kb of human genomic sequence and ~15 kb of chimpanzee genomic sequence. We developed models to explain the observed bimodality of the deletion size distribution, and showed that overall, *in vivo* deletions are smaller than those observed in cell culture analyses. Next, we quantified *Alu* recombination-mediated deletion in the human genome subsequent to the human-chimpanzee divergence and described 492 deletions (totaling ~400 kb of human genomic sequence) attributable to this process. Interestingly, the majority of these deletions are located within known or predicted genes, opening the possibility that a portion of the phenotypic differences

between humans and chimpanzees may be attributed to this mechanism. In the third analysis, we reported the *in vivo* existence of an endonuclease-independent insertion pathway for L1 elements and characterized twenty-one loci where L1 elements appear to have bridged genomic lesions. We show that these insertions are structurally distinguishable from classical L1 elements and suggest that this pathway may escape the purifying selection thought to be acting on endonuclease-dependent L1 loci in the genome.

CHAPTER ONE:

BACKGROUND

The rapid progress of high-throughput DNA sequencing technology has made whole-genome analysis almost a routine procedure (Shendure et al. 2004). In contrast to the $3 billion Human Genome Project, which published its data in 2001 (Lander et al. 2001b), the cost of sequencing a human genome in 2007 is less than $1 million (Check 2007), and it is expected that the expense and duration of such projects will continue to decrease drastically over the coming years (Dalton 2006). As publicly available sequence data accumulates exponentially, and increasingly accurate computational tools are developed to parse and annotate it, one of the most prominent discoveries has been that most of the DNA in any given genome appears to have no immediately discernable function (e.g., only ~1.4% of the human genome can unambiguously be shown to code for a definite protein) (IHGSC 2004; Lander et al. 2001b). Equally surprising, the large remainder of the genome, previously given such uncomplimentary epithets as "junk DNA", turns out to be part of a complex and dynamic network orchestrating the regulation and phenotypic expression of the tiny segment of coding DNA popularly referred to as "genes" (Brookfield 2005; Hedges and Batzer 2005; IHGSC 2004).

A majority of this "junk DNA" is comprised of transposable elements, which are pieces of genetic material that have the unique ability to move within the genome from one location to another (Britten and Kohne 1968; Lander et al. 2001b). First discovered by Barbara McClintock in the early 1950s during her work with color variegation in maize kernels (McClintock 1950; McClintock 1956), transposable elements have subsequently been found to be major components of genomes ranging from bacteria to humans (Campbell 2002; Deininger and Roy-Engel 2002). In mammals, transposable elements can be further subdivided into DNA transposons, which excise themselves completely from one genomic location before moving to another (Smit 1996) and retrotransposons, which multiply via a RNA intermediate, thus duplicating the original locus

each time they move (Deininger and Batzer 2002). As a consequence of this "copy-paste" mode of mobilization, the number of retrotransposons in any mammalian genome is usually orders of magnitude higher than that of DNA transposons, and they have a more prominent role in both global sequence architecture and genomic fluidity (Hattori et al. 2000; Lander et al. 2001b; Smit 1996).

In the context of the human genome, by far the two most successful families of retrotransposons are the LINE-1 and *Alu* elements, with copy numbers of ~520,000 and ~1.2 million, respectively (Lander et al. 2001b). The longer of these, LINE-1 (Long interspersed element-1 or L1) extends to about 6 kb in its full-length, functional form and comprises a 5' untranslated region (5'-UTR) bearing an internal RNA polymerase II promoter, followed by two non-overlapping open reading frames (ORF1 and ORF2, separated by a ~60 bp-long spacer), and a 3' UTR ending in a variable-length poly(A) tail (Kazazian and Moran 1998). The smaller ORF1 encodes a heterotrimeric RNA-binding protein that has nucleic acid chaperone activity *in vitro*, while the larger ORF2 encodes both reverse transcriptase and endonuclease activities. Interestingly, the LINE mRNA is transcribed bicistronically, in contrast to most mammalian RNAs (Feng et al. 1996; Kolosha and Martin 1997; Martin et al. 2003; Mathias et al. 1991). The shorter *Alu* element is ~300 bp long, transcribed by RNA polymerase III and ancestrally derived from the 7SL RNA gene (Kriegs et al. 2007). Each *Alu* element is a dimer-like structure; the 3' monomer has an additional 31 bp insertion relative to the 5' monomer, and is followed by a variable-length poly(A) tail (Deininger and Batzer 2002; Quentin 1992). *Alu* elements do not code for any enzymatic activity and instead parasitize upon the proteins synthesized during LINE transcription for their own retrotransposition, thus making them "parasite's parasites" (Schmid 2003).

Currently, it is thought that the mobilization of both *Alu* and L1 elements occurs through a mechanism termed target-site primed reverse transcription (TPRT) (Cost et al. 2002; Luan et al. 1993). During TPRT, the L1 endonuclease cleaves one strand of the host genomic DNA at a sequence loosely resembling 5'-TTTT/A-3' (where / denotes the cleavage site), producing a free 3'-hydroxyl (Cost and Boeke 1998; Feng et al. 1996). Next, the retrotransposon mRNA anneals to the nick site using its 3' poly (A) tail and the L1 reverse transcriptase synthesizes the retrotransposon cDNA using the mRNA as a template. Cleavage of the second DNA strand by the L1 endonuclease usually takes place 7-20 base pairs downstream of the initial nicking site, creating staggered breaks in the genomic DNA that are later filled in to form direct repeats flanking the newly inserted element (termed target site duplications or TSDs)(Szak et al. 2002). Integration of the newly synthesized cDNA and completion of second-strand synthesis are the remaining steps in the TPRT model; however, the order in which they occur and their exact mechanism are only beginning to be elucidated. Template jumping of the L1 reverse transcriptase from the 5'-end of the newly synthesized cDNA to the host genome and small stretches of complementary base pairing at the 5' junction of the integration complex are thought to play a vital part in completing the TPRT process (Babushok and Kazazian 2007; Martin et al. 2005; Zingler et al. 2005).

Although the finished sequence of the human genome demonstrates that almost half our DNA is composed of mobile elements and it is evident that at least two families of such elements (*Alu* and LINE) have been actively mobilizing in the recent evolution of the human lineage (i.e., after its divergence from a common ancestor with the chimpanzee)(Boissinot et al. 2000; Carter et al. 2004; Myers et al. 2002; Otieno et al. 2004; Salem et al. 2003), a myriad questions about the biology of these elements remains unanswered. The more popular opinion in the literature is

that on a global scale, both these families are most likely deleterious and at best neutral within the genome, and have achieved their high numbers through a finely tuned strategy of parasitism (Boissinot et al. 2001; Cordaux et al. 2006b; Schmid 2003). However, at counterpoint to this school of thought are the various analyses that have proposed different "functional" roles for *Alu* and LINE elements such as origins of replication, gene expression regulators, agents of DNA repair and X-chromosome inactivation or scaffolds for meiotic replication (Brosius and Gould 1992; Liu et al. 1995; Schmid 1998). From an objective viewpoint, both these schools need not be reciprocally exclusive, and it may be overly simplistic to treat the interactions between the *Alu* and L1 families and primate genomes as being a zero-sum game. Indeed, a systems biological approach wherein the genome and these elements are seen in the context of an ecosystem, may be a suitable way of representing this complex relationship (Brookfield 2005). As such, it is likely that the broad question will still remain open for some time as to whether retrotransposons serve any purpose at all within the human genome, or if their historical characterization as "selfish DNA" entities is justified.

Within this context, the remarkably high copy numbers of the Alu and LINE families in the human genome (~1.2 million and ~520,000, respectively) effectively makes them ubiquitous stretches of non-allelic sequence homology distributed over the length of all chromosomes, uniquely predisposing them to participation in genomic rearrangements. Structurally, apart from insertional mutagenesis during retrotransposition, both *Alu* and LINE elements can induce other forms of change in local sequence architecture, including but not limited to recombination-mediated deletions, insertion-mediated deletions, segmental duplications, inversions and inter-chromosomal and intra-chromosomal translocations of host genomic sequence (Hedges and Deininger 2007). Additionally, it is now evident that both *Alu* and LINE mRNA are occasionally

co-opted by genomic DNA repair machinery to serve as molecular Band-Aids® for repairing potentially lethal double-strand breaks (Morrish et al. 2002; Sen et al. 2007). From a functional viewpoint, *Alu* elements have been "exonized" at a number of loci in the human genome, can cause alternative splicing and show transcriptional responses to cellular stress that downregulate translational activity (Dagan et al. 2004; Liu et al. 1995; Sorek et al. 2002; Sorek et al. 2004), while LINE elements have been associated with gene-breaking, exon shuffling and transcriptional control of gene expression (Babushok et al. 2007; Matlik et al. 2006; Moran et al. 1999; Speek 2001; Wheelan et al. 2005). However, it is important to remember that only those genomic rearrangements occurring in germline cells and transmitted to succeeding generations (Han et al. 2005; Sen et al. 2006; Sen et al. 2007) hold any long-term evolutionary significance, given that somatic retrotransposon-mediated instability with strong deleterious effects (e.g., the wide variety of diseases associated with the Alu and LINE families; reviewed in (Deininger and Batzer 1999) would attract immediate and strong negative selection and be restricted to a blip on the genomic radar screen. Hence, the true impact of these elements on human genome stability over evolutionary timescales can only be ascertained through comparative genomic analyses of our genome with closely related primate genomes (Ebersberger et al. 2002; Han et al. in press; Han et al. 2005; Lee et al. 2007; Sen et al. 2006). The recent availability of complete sequences for the chimpanzee and rhesus macaque genomes (CSAC 2005; RMGSAC 2007) has facilitated the development of computational tools to accurately analyze human lineage-specific changes in local sequence architecture, leading to an acceleration of research in this field (Disotell and Tosi 2007; Varki and Altheide 2005). My dissertation is part of this recent surge, and the work which follows focuses primarily on different aspects of structural dynamics of LINE and Alu elements within the human genome.

In chapter two, we describe 50 genomic deletions directly linked to the insertion of L1 elements, resulting in the loss of ~18 Kb of sequence from the human genome and ~15 Kb from the chimpanzee genome. Our data suggest that during the primate radiation, L1 insertions may have deleted up to 7.5 Mb of target genomic sequences. We report a pattern of genomic deletion sizes similar to those created during the retrotransposition of *Alu* elements (Callinan et al. 2005). This analysis supports the existence of different mechanisms for small and large L1 insertion-mediated deletions. We present a model for the correlation of L1 insertion size and the corresponding deletion size, and show that internal rearrangements can modify L1 structure during retrotransposition events associated with large deletions. While the results of our *in vivo* analysis appear to conflict with previous cell culture assays of L1 insertion-mediated deletions (Gilbert et al. 2002; Symer et al. 2002) in terms of the size and rate of sequence deletion, evolutionary factors can reconcile the differences.

In chapter three, we compare the human and chimpanzee genomes to determine the magnitude of *Alu* recombination-mediated deletion in the human genome since the human-chimpanzee divergence ~5-7 million years ago (Chen and Li 2001). Combining computational data mining and experimental verification techniques, we identified 492 human-specific deletions (totaling ~400 Kb) attributable to this process, making it a significant component of the insertion/deletion spectrum of the human genome. The majority of the deletions (295/492) coincide with known or predicted genes (including three that deleted functional exons as compared to orthologous chimpanzee genes), implicating this process in creating a substantial portion of the genomic differences between humans and chimpanzees. Overall, we find that *Alu* recombination-mediated genomic deletion has had a much higher impact than that reflected by

previously identified isolated events, and that it continues to contribute to the dynamic nature of the human genome.

In chapter four, we analyzed the human genome to demonstrate that an alternative, endonuclease-independent pathway for L1 insertion that was hitherto known only in DNA-repair deficient cell lines (Morrish et al. 2002) has also been active in recent human genome evolution. We characterized twenty-one loci where L1 elements have integrated without signs of endonuclease related activity. The structural features of these loci suggest a role for this process in DNA double-strand break repair. We show that endonuclease-independent L1 insertions are structurally distinguishable from classical L1 insertion loci and that they are associated with inter-chromosomal translocations and deletions of target genomic DNA.

References

Babushok, D.V. and H.H. Kazazian, Jr. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28:** 527-539.

Babushok, D.V., K. Ohshima, E.M. Ostertag, X. Chen, Y. Wang, P.K. Mandal, N. Okada, C.S. Abrams, and H.H. Kazazian, Jr. 2007. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Res* **17:** 1129-1138.

Boissinot, S., P. Chevret, and A.V. Furano. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17:** 915-928.

Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18:** 926-935.

Britten, R.J. and D.E. Kohne. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161:** 529-540.

Brookfield, J.F. 2005. The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet* **6:** 128-136.

Brosius, J. and S.J. Gould. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci U S A* **89:** 10706-10710.

Callinan, P.A., J. Wang, S.W. Herke, R.K. Garber, P. Liang, and M.A. Batzer. 2005. Alu Retrotransposition-mediated Deletion. *J Mol Biol* **348:** 791-800.

Campbell, A. 2002. Eubacterial Genomes. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 1024-1039. ASM Press, Washington, D.C.

Carter, A.B., A.H. Salem, D.J. Hedges, C.N. Keegan, B. Kimball, J.A. Walker, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2004. Genome-wide analysis of the human Alu Yb-lineage. *Hum Genomics* **1:** 167-178.

Check, E. 2007. James Watson's genome sequenced. In *Nature News*. Nature Publishing Group.

Chen, F.C. and W.H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444-456.

Cordaux, R., J. Lee, L. Dinoso, and M.A. Batzer. 2006. Recently integrated *Alu* retrotransposons are essentially neutral residents of the human genome. *Gene* **373:** 138-144.

Cost, G.J. and J.D. Boeke. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37:** 18081-18093.

Cost, G.J., Q. Feng, A. Jacquier, and J.D. Boeke. 2002. Human L1 element target-primed reverse transcription in vitro. *Embo J* **21:** 5899-5910.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87.

Dagan, T., R. Sorek, E. Sharon, G. Ast, and D. Graur. 2004. AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res* **32:** D489-492.

Dalton, R. 2006. Sequencers step up to the speed challenge. *Nature* **443:** 258-259.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67:** 183-193.

Deininger, P.L. and M.A. Batzer. 2002. Mammalian retroelements. *Genome Res* **12:** 1455-1465.

Deininger, P.L. and A.M. Roy-Engel. 2002. Mobile Elements in Animal and Plant Genomes. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 1074-1092. ASM Press, Washington, D.C.

Disotell, T.R. and A.J. Tosi. 2007. The monkey's perspective. *Genome Biol* **8:** 226.

Ebersberger, I., D. Metzler, C. Schwarz, and S. Paabo. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* **70:** 1490-1497.

Feng, Q., J.V. Moran, H.H. Kazazian, Jr., and J.D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905-916.

Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Han, K., J. Lee, T.J. Meyer, J. Wang, S.K. Sen, D. Srikanta, P. Liang, and M.A. Batzer. in press. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genetics*.

Han, K., S.K. Sen, J. Wang, P.A. Callinan, J. Lee, R. Cordaux, P. Liang, and M.A. Batzer. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* **33:** 4040-4052.

Hattori, M., A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311-319.

Hedges, D.J. and M.A. Batzer. 2005. From the margins of the genome: mobile elements shape primate evolution. *Bioessays* **27:** 785-794.

Hedges, D.J. and P.L. Deininger. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* **616:** 46-59.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Kazazian, H.H., Jr. and J.V. Moran. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet* **19:** 19-24.

Kolosha, V.O. and S.L. Martin. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci U S A* **94:** 10155-10160.

Kriegs, J.O., G. Churakov, J. Jurka, J. Brosius, and J. Schmitz. 2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* **23:** 158-161.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lee, J., R. Cordaux, K. Han, J. Wang, D.J. Hedges, P. Liang, and M.A. Batzer. 2007. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390:** 18-27.

Liu, W.M., W.M. Chu, P.V. Choudary, and C.W. Schmid. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23:** 1758-1765.

Luan, D.D., M.H. Korman, J.L. Jakubczak, and T.H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72:** 595-605.

Martin, S.L., D. Branciforte, D. Keller, and D.L. Bain. 2003. Trimeric structure for an essential protein in L1 retrotransposition. *Proc Natl Acad Sci U S A* **100:** 13815-13820.

Martin, S.L., W.L. Li, A.V. Furano, and S. Boissinot. 2005. The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenet Genome Res* **110:** 223-228.

Mathias, S.L., A.F. Scott, H.H. Kazazian, Jr., J.D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808-1810.

Matlik, K., K. Redik, and M. Speek. 2006. L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **2006:** 71753.

McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36:** 344-355.

McClintock, B. 1956. Intranuclear systems controlling gene action and mutation. *Brookhaven Symp Biol***:** 58-74.

Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530-1534.

Morrish, T.A., N. Gilbert, J.S. Myers, B.J. Vincent, T.D. Stamato, G.E. Taccioli, M.A. Batzer, and J.V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31:** 159-165.

Myers, J.S., B.J. Vincent, H. Udall, W.S. Watkins, T.A. Morrish, G.E. Kilroy, G.D. Swergold, J. Henke, L. Henke, J.V. Moran et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71:** 312-326.

Otieno, A.C., A.B. Carter, D.J. Hedges, J.A. Walker, D.A. Ray, R.K. Garber, B.A. Anders, N. Stoilova, M.E. Laborde, J.D. Fowlkes et al. 2004. Analysis of the Human Alu Ya-lineage. *J Mol Biol* **342:** 109-118.

Quentin, Y. 1992. Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. *Nucleic Acids Res* **20:** 487-493.

Rhesus Macaque Genome Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222-234.

Salem, A.H., J.S. Myers, A.C. Otieno, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003. LINE-1 preTa elements in the human genome. *J Mol Biol* **326:** 1127-1146.

Schmid, C.W. 1998. Does SINE evolution preclude Alu function? *Nucleic Acids Res* **26:** 4541-4550.

Schmid, C.W. 2003. Alu: a parasite's parasite? *Nat Genet* **35:** 15-16.

Sen, S.K., K. Han, J. Wang, J. Lee, H. Wang, P.A. Callinan, M. Dyer, R. Cordaux, P. Liang, and M.A. Batzer. 2006. Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* **79:** 41-53.

Sen, S.K., C.T. Huang, K. Han, and M.A. Batzer. 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35:** 3741-3751.

Shendure, J., R.D. Mitra, C. Varma, and G.M. Church. 2004. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5:** 335-344.

Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* **6:** 743-748.

Sorek, R., G. Ast, and D. Graur. 2002. Alu-containing exons are alternatively spliced. *Genome Res* **12:** 1060-1067.

Sorek, R., G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur, and G. Ast. 2004. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell* **14:** 221-231.

Speek, M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol* **21:** 1973-1985.

Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Szak, S.T., O.K. Pickeral, W. Makalowski, M.S. Boguski, D. Landsman, and J.D. Boeke. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052.

Varki, A. and T.K. Altheide. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res* **15:** 1746-1758.

Wheelan, S.J., Y. Aizawa, J.S. Han, and J.D. Boeke. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15:** 1073-1078.

Zingler, N., U. Willhoeft, H.P. Brose, V. Schoder, T. Jahns, K.M. Hanschmann, T.A. Morrish, J. Lower, and G.G. Schumann. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15:** 780-789.

CHAPTER TWO:

GENOMIC REARRANGEMENTS BY LINE-1 INSERTION-MEDIATED DELETION IN THE HUMAN AND CHIMPANZEE LINEAGES*

*Reprinted by permission of Nucleic Acids Research

Introduction

Long INterspersed Elements (LINE-1s or L1s) are abundant non-LTR retrotransposons in mammalian genomes and comprise ~17% of the human genome (Lander et al. 2001). They have reached copy numbers of about 520,000 (Lander et al. 2001; Ostertag and Kazazian 2001a) and have expanded over the past 100-150 million years (Smit *et al.* 1995). In their full-length state, they are capable of autonomous retrotransposition through an RNA intermediate. However, ~ 99.8% of extant L1s in the human genome are retrotransposition-defective (Sassaman *et al.* 1997), either due to point mutations or larger changes such as 5' truncations, 5' inversions or other internal rearrangements (Gilbert et al. 2002; Kazazian and Moran 1998; Myers et al. 2002; Ostertag and Kazazian 2001b). While extant human L1-derived elements have an average size of 900 bp for all L1 copies (Lander et al. 2001), an active full-length L1 element is about 6 Kb in length, and encodes two open reading frames (ORFs) separated by a 63 bp spacer region. The first L1-encoded protein, ORF1p, is a 40 kDa RNA-binding protein, while the second, ORF2p, is a 150 kDa protein with both endonuclease (EN) and reverse transcriptase (RT) activities (Feng et al. 1996; Mathias et al. 1991). The two ORFs are preceded by a 5' untranslated region (5'-UTR), which contains an internal promoter for RNA polymerase II, and are followed by a 3' UTR ending in a poly(A) tail. The L1-encoded proteins predominantly exhibit *cis*-preference, transposing the same RNA that encoded them (Dewannieux et al. 2003; Wei et al. 2001).

The number of full-length retrotransposition-competent L1 elements that are currently estimated to be propagating in the human genome, however, is much lower than the total number of insertions, with estimates varying between 60 and 100 elements (Brouha et al. 2003; Kazazian and Goodier 2002; Sassaman et al. 1997). The mobilization of L1 elements is based on a mechanism termed target-primed reverse transcription (TPRT) which provides useful landmarks

for the identification of L1 insertion (Luan *et al.* 1993). During this process, a single-strand nick in the genomic DNA is made by the L1 EN at the 5'-TTTT/A-3' consensus cleavage site (Cost and Boeke 1998; Feng et al. 1996; Jurka 1997; Morrish et al. 2002) on the antisense strand, after which the L1 RNA transcript anneals by its poly(A) tail to the cleavage site and primes reverse transcription. After the synthesis of the complementary DNA copy and its covalent attachment to the target DNA, second strand synthesis occurs using the first strand as a template. Single-stranded regions remaining in the target DNA at either end are filled in to create target site duplications (TSDs), structural hallmarks of the TPRT process which have been used in the computational location of L1 insertions (Szak *et al.* 2002). However, in situations where L1 integration results in the deletion of portions of target DNA, TSDs may not be formed, and a number of studies have reported L1 insertions without TSDs of any length (Gilbert et al. 2005; Morrish et al. 2002).

Both mammalian cell culture assays and previous genomic analyses have implicated L1s as agents in complex genomic rearrangements. Mechanisms of L1-mediated genomic instability include (i) unequal homologous recombination between L1 elements (Burwinkel and Kilimann 1998; Ostertag and Kazazian 2001a); (ii) generation of interstitial (> 3 Kb) deletions in the target sequence (Gilbert *et al.* 2002; Symer *et al.* 2002) and (iii) transduction of varying amounts of 3' flanking sequence along with the L1 itself during retrotransposition (Pickeral *et al.* 2000). The last process is also a mechanism for L1-mediated exon shuffling (Goodier et al. 2000; Moran et al. 1999; Pickeral et al. 2000). The L1 enzymatic machinery may also be utilized during pseudogene processing and *Alu* element mobilization (Dewannieux et al. 2003; Wei et al. 2001).

Previous analyses of genomic deletions created upon L1 retrotransposition in human DNA have almost exclusively relied on cell culture assays and described *de novo* L1 retrotransposition events associated with target site deletions (Gilbert *et al.* 2002; Symer *et al.*

2002). Large interstitial deletions, ranging up to 71 Kb, have been reported as one of the consequences of L1 retrotransposition (Gilbert *et al.* 2002). However, the artificially constructed L1 insertion cassettes utilized in these assays permit the recovery of large and full-length L1 insertions only, and the extent of genomic deletion identified in these analyses may not represent the actual extent of existing deletions associated with L1 insertions in the human genome. The recent completion of the draft chimpanzee genome sequence (PanTro1; Nov. 2003 freeze) provides the first opportunity to locate and quantify in an evolutionary framework existing human-specific and chimpanzee-specific L1 insertion-mediated deletions (L1IMDs). In this study, we identified species-specific L1IMD candidates via computational screening of the draft genomic sequences of *Homo sapiens* and *Pan troglodytes* and confirmed them experimentally. We find that L1 insertions are directly responsible for the removal of ~18 Kb of human genomic sequence and ~15 Kb of chimpanzee genomic sequence within the past 4-6 million years and may have generated over 11,000 deletion events during the radiation of the primate order, resulting in the removal of up to 7.5 Mb of DNA in the process. We also propose mechanisms to explain the correlation of L1 insertion size with the size of the deletion it causes and suggest models for the formation of truncation/inversion structures during L1 integration processes associated with target site deletions.

Results

A Genome-Wide Analysis of Human- and Chimpanzee-Specific L1IMDs

To locate L1IMD loci in the human and common chimpanzee lineages, we first compared data from the draft human and common chimpanzee genomic sequences. We computationally detected 30 human-specific and 33 chimpanzee-specific L1 insertion candidates associated with extra (non-homologous) sequences at the orthologous loci in the other genome. PCR display and

manual inspection of the DNA sequences resulted in the exclusion of four human loci and six chimpanzee loci as false positives for L1IMD. These cases were due to stretches of Ns in the chimpanzee genome assembly (corresponding to unsequenced regions) or species-specific *Alu* element insertions in the 5' end of the loci, leading to partial mismatches at the orthologous locus in the other species, one of the prerequisites in our computational approach to identify candidate L1IMD loci. This resulted in the validation of 26 and 27 L1IMDs identified from the human and chimpanzee genomes, respectively. PCR analysis of all but one (LH4) L1IMD loci in five primate species showed that all the L1IMDs were specific to the species from which they were identified (Figure 2.1). Locus LH4 could not be amplified due to the presence of other repeat elements in the flanking sequence. However, on the basis of (i) the 99.5% similarity of the L1 element inserted at this locus to the consensus sequence of the human-specific L1Hs subfamily,

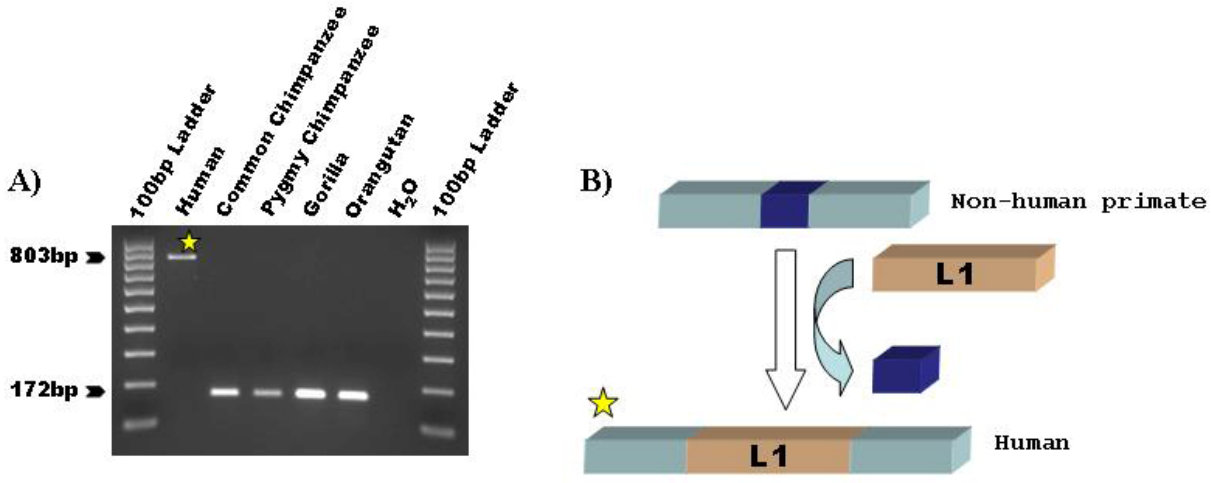


Figure 2.1. L1 insertion-mediated deletion in the human genome. (A) Gel chromatographs of PCR products from a phylogenetic analysis of the human-specific L1IMD are shown. The DNA template used in each lane is shown at top. The product sizes for filled and empty alleles are indicated at the left. (B) The schematic diagrams depict the insertion of the L1 element (orange boxes) and the deletion of genomic DNA (blue boxes). Flanking unique DNA sequences are shown as light blue boxes.

and (ii) the presence of extra (non-homologous) genomic sequence at this locus in the common chimpanzee genome, the L1 insertion and associated deletion at locus LH4 were included in our dataset of human-specific genomic deletions directly associated with L1 insertion.

Because the L1 elements associated with L1IMD were not flanked by TSDs, the only possible hallmark of TPRT in our L1IMD events was the presence of L1 EN cleavage sites. To confirm that the deletions observed in the human and chimpanzee genomes were generated during the process of L1 insertion rather than prior to (and therefore independently of) the L1 insertion, we looked for L1 EN cleavage motifs in our L1IMD loci and divided the loci into categories based on the number of differences with the 5'-TTTT/A-3' consensus L1 EN cleavage site (Boeke and Devine 1998; Jurka 1997; Morrish et al. 2002). For each locus, we compared the sequence corresponding to the insertion site predicted to the consensus EN cleavage motif to see if it was L1 EN-generated or not. To conservatively exclude 'false' cleavage motifs arising from post-insertion mutations mimicking the L1 EN consensus cleavage sequence, we down-weighted the number of transition differences with the consensus EN cleavage motif by a factor 0.5 because transitions in the cleavage site that conserve the homopurine or homopyrimidine runs are generally better tolerated by the EN than transversions (Cost *et al.* 2001). Additionally, we further down-weighted transitions by a second factor 0.5, because of their more frequent occurrence than transversions in GC-poor regions (Nachman and Crowell 2000). In both humans and chimpanzees, the frequency spectra of the integration site preferences showed unimodal distributions with modes at 0.5 differences from the consensus sequence 5'-TTTT/A-3' (Figure 2.2). The L1 EN site preference of our L1IMDs is thus very similar to that of L1-Ta subfamily elements (n = 282) identified in a previous study (Morrish *et al.* 2002). However, three of the 53 loci (LH11, LH12 and LC6) identified computationally as L1IMD candidates had cleavage sites

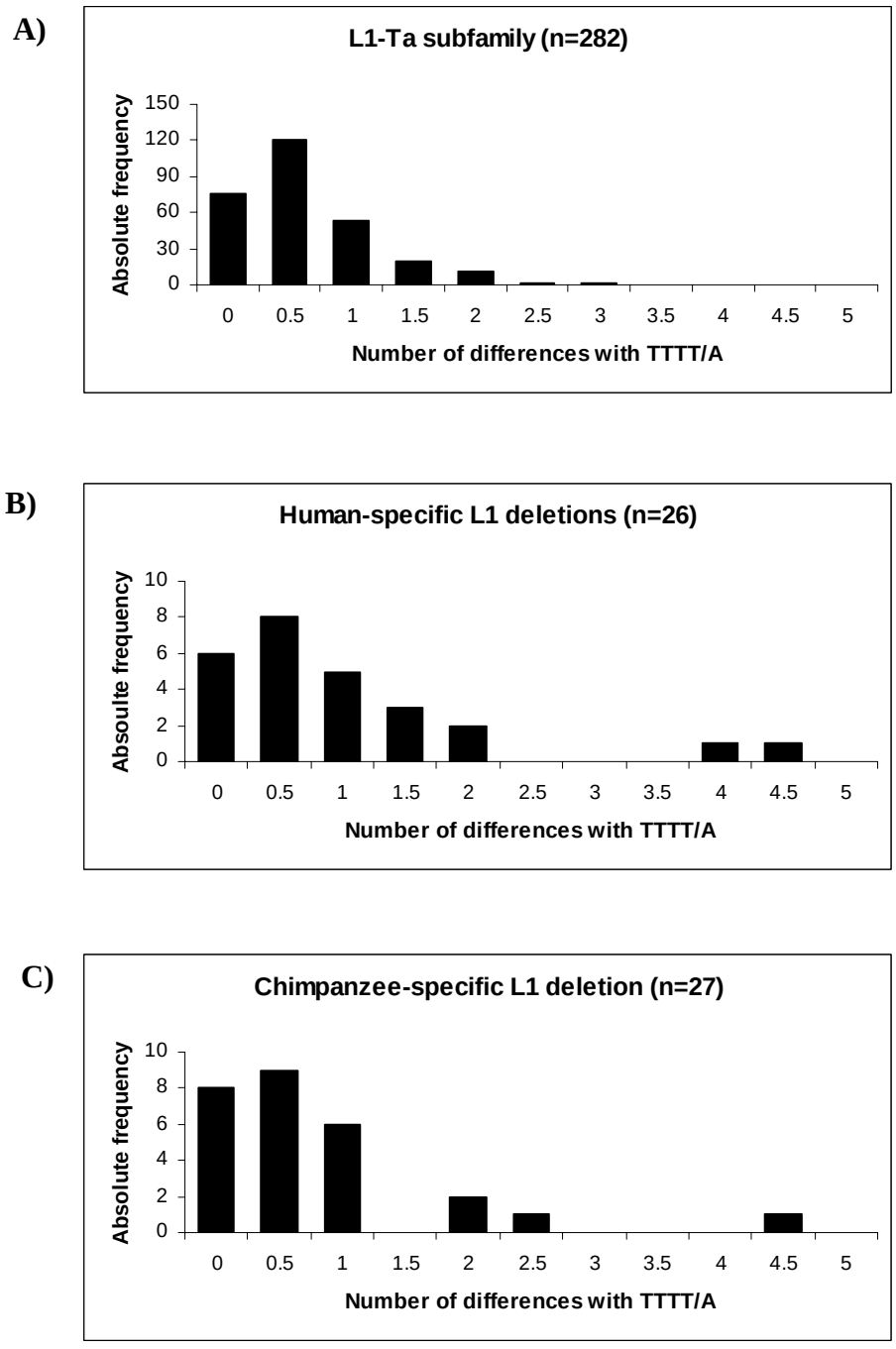


Figure 2.2. Endonuclease cleavage site preferences for the L1IMDs. The number of differences from the consensus L1 endonuclease cleavage site (TTTT/A) are shown after down-weighting transitions. The data are analyzed for (A) The L1-Ta subfamily elements identified in Morrish et al. (2002); (B) Human lineage specific L1 insertions (LH11 and LH12 excluded as number of differences ≥2.5); (C) Chimpanzee lineage specific L1 insertions (LC6 excluded as number of differences ≥2.5)

substantially differing from the consensus by four or more substitutions while the maximum number of substitutions observed in the L1-Ta subfamily is three (Figure 2.2), hence casting doubt on the use of EN during insertion of these elements. We believe that these deletions are the products of EN independent insertions similar to those reported in previous cell culture assays (Morrish *et al.* 2002). To be conservative, these three elements were removed from the analyses, resulting in a final dataset of 24 and 26 L1IMD loci in the human and chimpanzee genomes, respectively, with deletions produced unambiguously by an L1 EN-dependent mechanism.

Characteristics of the L1 Insertions Associated with L1IMDs

The L1 insertions in our study ranged in size from 61 to 5174 bp. Of the 24 human L1 insertions, eight belonged to the L1Hs subfamily according to RepeatMasker, 14 to L1PA2 and 2 could not be confidently assigned to any subfamily. As to the 26 chimpanzee L1 insertions, 23 belonged to the L1PA2 subfamily, one to L1PA5 while two could not be confidently assigned to any subfamily. Median-joining network analysis (Figure 2.3) of the L1 elements in our study, using substitutions at the 4 key subfamily-diagnostic sequence positions (i.e., bp 5930-5932 and 6015 in the 3' UTR of the full-length L1 consensus sequence) shows that the chronological order in evolutionary time (from youngest to oldest) of the L1 elements in our study is Ta (ACA/G) - PreTa (ACG/G) - ACG/A - GCG/A or AAG/A- GCG/G - L1PA2 (GAG/A). This evolutionary order is consistent with previous analyses of L1 insertions utilizing other phylogenetic approaches such as neighbor-joining, maximum-likelihood and maximum parsimony analyses (Ovchinnikov *et al.* 2002).

All the elements were 5' truncated to different degrees (Ovchinnikov *et al.* 2001), with most having their 5' start position located in the 3' UTR of the consensus full-length L1.3 reference sequence (Dombroski *et al.* 1993) (Table 2.1).The size distribution of the L1 insertions

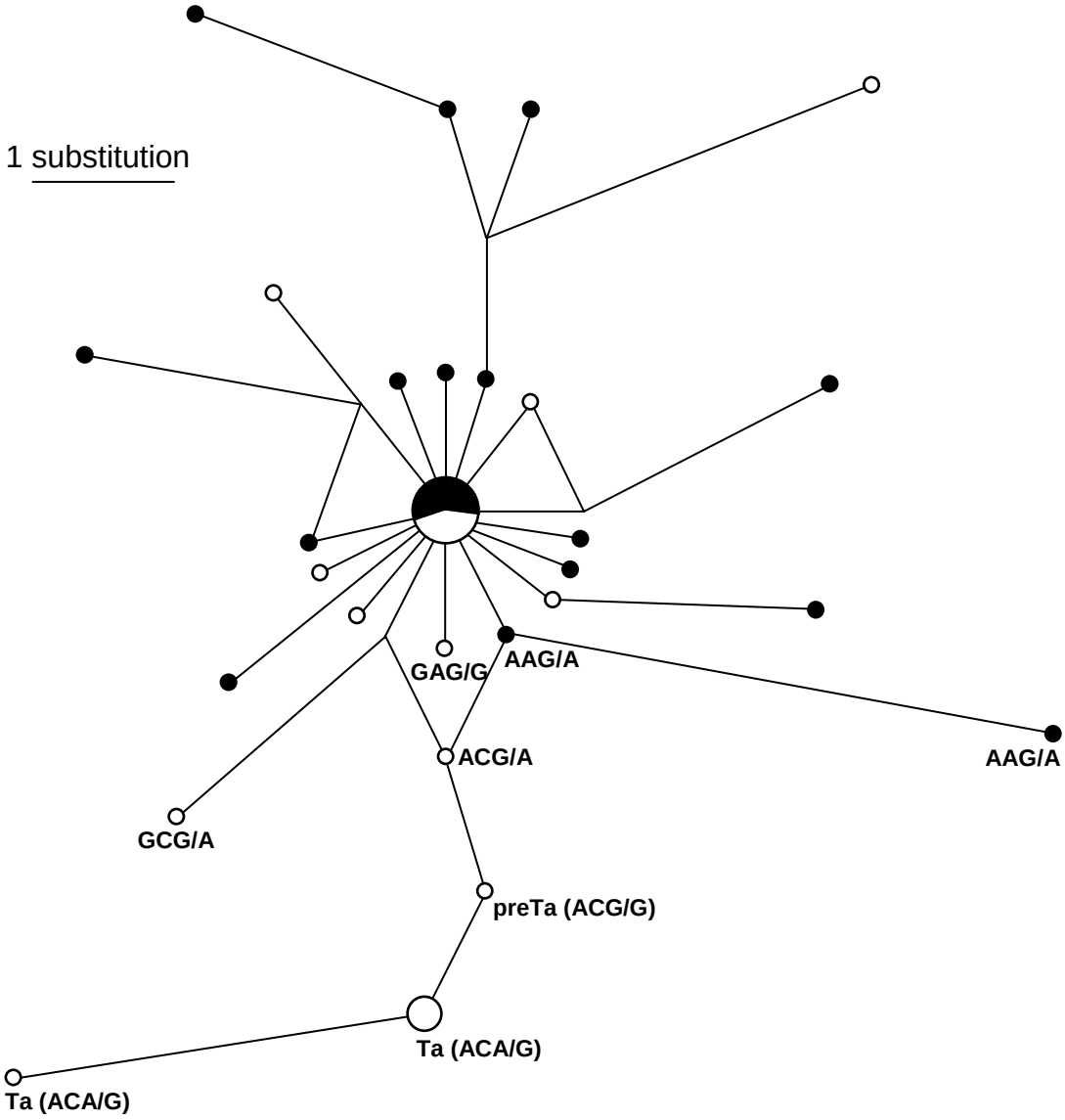


Figure 2.3. Median-joining network of the L1 elements associated with L1IMD. Empty circles denote human-specific L1 elements. Filled circles denote chimpanzee-specific L1 elements. The size of circles indicates the number of L1 loci with that sequence type. The lines denote substitution steps, with a one-step distance indicated in the top-left corner. The subfamily-specific diagnostic sequence positions (corresponding to positions 5930-5932 and 6015 in the 3' UTR of the full-length L1 consensus sequence) are specified below each relevant node.

Table 2.1. Structural summary of L1 insertion-mediated deletions

Feature	Human	Chimpanzee
Full-length L1 insertions	0	0
5' truncated L1 insertions	24	26
Internal rearrangements	4	2
Non-inverted	4	0
5'truncation/inversions	0	2
With TSDs of any length	0	0
Total L1 size (bp)	31,617	25,031
Mean of L1 size (bp)	1322	963
Total deletion size (bp)	17,671	14,923
Mean of deletion size (bp)	736	574
Median of deletion size (bp)	21	73

is similar to that obtained in a previous human cell culture assay of L1-mediated genomic instability (Symer *et al.* 2002). As to chromosomal distribution, the majority of the L1IMDs were located on chromosomes 1 to 12, which probably relates to both the larger size of these chromosomes and their higher density of truncated (3' intact) L1 insertions (Szak *et al.* 2002).

Four human-specific L1 insertions (at loci LH17, LH19, LH26 and LH31) showed the presence of partially duplicated or internally rearranged L1 segments, suggesting either an atypical structure for the particular L1 insertion or two independent L1 insertions into the same locus during a relatively short time. Given the size of the human genome (~3300 Mb), two L1 insertions occurring at exactly the same location four times in 24 human loci is very improbable

considering that there have been no instances of L1 element insertion homoplasy ever reported (Ho et al. 2005; Salem et al. 2003a; Salem et al. 2005). Loci LH17 and LH 31 each consist of two L1PA2 segments in the same orientation with 300 bp and 286 bp gaps between the two segments, respectively, relative to the L1PA2 consensus sequence. These loci probably represent single L1 insertion events associated with internal deletions. The other two loci, LH19 and LH26, each consist of two identical L1PA2 segments in tandem, with 53 bp and 189 bp stretches respectively being repeated in the same orientation without any intervening region. Two chimpanzee loci (LC26 and LC27) also presumably resulted from 5' truncation/inversion events, with overlapping junctions between the inverted segments (Szak *et al.* 2002).

The poly(A) tails of the L1 inserts ranged in length from 2 to 64 bases, with similar averages of 19 bases in humans and 21 bases in chimpanzees. Our value for the average poly(A) tail lengths for human L1 insertions is thus much lower than those from two previous cell culture assays of *de novo* L1 retrotransposition in HeLa cells, that reported averages of ~60 residues (Gilbert *et al.* 2002) and 88±27 residues (Symer *et al.* 2002). Furthermore, the 23 bp average length of the poly(A) tail among members of the youngest L1Hs subfamily was slightly higher than the 16 bp average for the older L1PA2 subfamily elements. Our data thus suggest the occurrence of post-insertional shortening of poly(A) tails over time, possibly due to replication slippage (Ovchinnikov *et al.* 2001; Roy-Engel *et al.* 2002). While the poly(A) tails in the *de novo* insertions identified in the aforementioned studies are exclusive runs of adenosine residues, the tails of the L1s identified in our study show considerable patterning and incidence of other nucleotide residues, with $TA_{(n)}$ being the most common pattern (six cases in the chimpanzee L1s and four cases in human L1s), which corroborates the findings of Szak et al. (2002). We found no significant correlation between the size of the poly(A) tail and the size of the L1 insertion in our dataset ($r = 0.12$, $P = 0.84$).

Characteristics of the L1IMDs

L1IMD events resulted in the deletion of 17,671 nucleotides from the human genome and 14,921 nucleotides from the chimpanzee genome (Table 2.1). The size distribution of the deletions (Figure 2.4) showed a strong bias towards the smaller sizes, with 50% of the chimpanzee L1IMDs and 58% of the human L1IMDs showing sizes of <200 bp. However, both human and chimpanzee events were also characterized by 20-30% of L1IMDs longer than 1 Kb. These observations were further reflected by the medians of the L1IMD sizes being an order of magnitude smaller than the average L1IMD size in both human and chimpanzee (Table 2.1). The L1IMD loci in our study in both human and chimpanzee lineages showed significant ($P < 0.05$ in both species) positive correlations between the size of the L1IMD and the size of its associated L1 insertion.

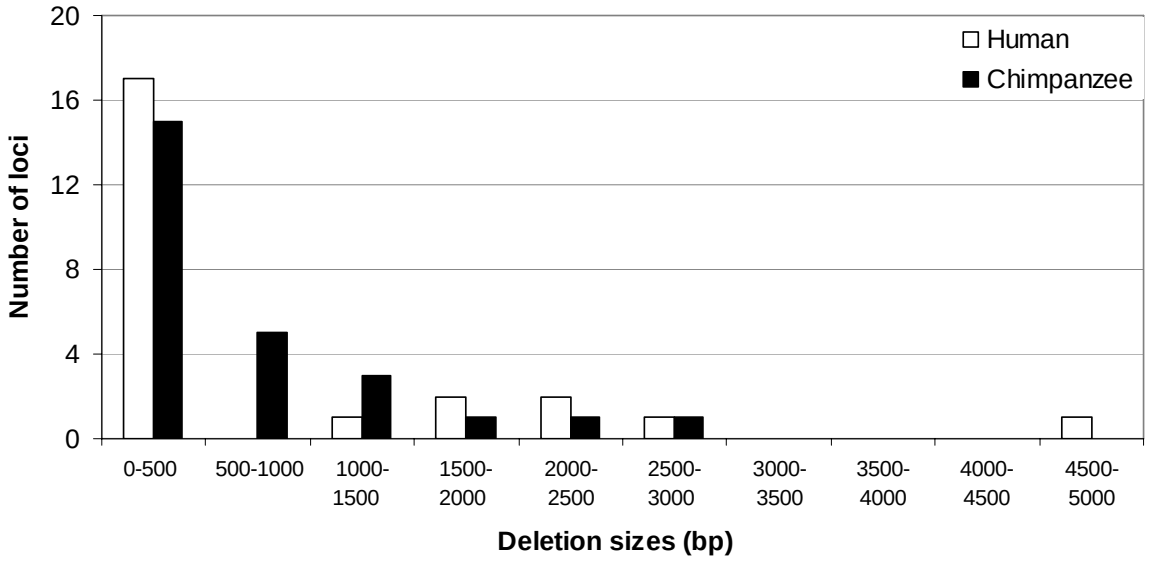


Figure 2.4. Size distribution of the L1IMDs. The size distribution of all the L1IMD events identified in the human and chimpanzee lineages is displayed in 500bp intervals or bins.

L1IMD Polymorphism

To estimate the level of polymorphism associated with human-specific L1IMD loci, we amplified them in 80 individuals from four geographically diverse populations. In all, five out of 23 loci (~22%) were polymorphic (Table 2.2), three of which contained L1Hs elements and two contained L1PA2 elements. Within our common chimpanzee panel of 12 individuals, four out of 26 loci (~15%) were polymorphic (Table 2.2), three of which contained L1PA2 elements and one contained a L1PA5 element. Overall, this indicates that human L1IMDs are associated with slightly higher polymorphism rates than their chimpanzee counterparts. These results contrast with those obtained for *Alu* retrotransposition-mediated deletions (ARDs) (Callinan *et al.* 2005) and *Alu* insertions (Hedges *et al.* 2004) in the context of human/chimpanzee comparisons, in which the polymorphism rates were found to be about twice as high in chimpanzee as in human. These data could be indicative of a slowdown of L1 retrotransposition within the chimpanzee lineage as compared to the human lineage.

Table 2.2. L1 insertion-mediated deletion frequency and polymorphism levels within the human and chimpanzee lineages

	Human	Chimpanzee	Human to Chimpanzee ratio
Total observed L1IMDs	24	26	0.92
PCR amplified	23	26	-
Fixed present	18	22	-
Polymorphic loci	5	4	-
Polymorphic fraction	0.22	0.15	1.41
Adjusted polymorphic loci	10	8	
Adjusted number of L1IMDs	29	30	

Genomic Environment of L1IMDs

Contrary to non-autonomous *Alu* elements, L1s seem to have a preference for GC-poor regions of the genome (Boissinot et al. 2004; Ovchinnikov et al. 2001), which may be a consequence of either the L1 EN site preference (Cost and Boeke 1998) or of faster removal of L1s from GC–rich regions (Boissinot *et al.* 2001). To analyze whether L1 insertions causing deletions in the target sequence behaved differently from typical insertions, we analyzed GC content of 40 Kb of the flanking sequences (20 Kb each from the 5' and 3' ends) of the L1IMDs. Because poly(A) tails are shortened over time by the combined effects of mutation and replication slippage (Ovchinnikov *et al.* 2001) causing the presence of 'fossil' poly(A) tails in the 3' flanking sequence, we avoided bias towards excessive adenosine residues by counting 20 Kb at the 3' end after excluding 100 bp from the end of the poly-adenylation signal (AATAAA) of the L1 inserts. The mean GC content for the flanking regions of the human-specific and chimpanzee-specific L1IMDs was 38% and 39%, respectively. Compared to the ~42% average GC content of the draft human and chimpanzee genomes (Lander et al. 2001; Watanabe et al. 2004), L1IMD loci thus seem to be concentrated in AT-rich areas of the genome. Remarkably, ARDs in the human and chimpanzee genomes also show a preference for AT-rich locations (Callinan *et al.* 2005). The reduced GC content (~36%) around the eight youngest human L1 elements belonging to the L1Hs subfamily in our dataset (LH4, LH15, LH17, LH19, LH20, LH22, LH23, LH24) is consistent with previous findings (Boissinot *et al.* 2004).

To further characterize the genomic context in which L1IMDs occur, we calculated known and predicted gene densities in 4, 2 and 0.5 Mb windows lying immediately 5' and 3' to the L1IMDs (see supplementary data for gene counts in Batzer Laboratory Web site). Our results indicate that L1IMDs are concentrated in regions of low gene density (*i.e.* 1 gene per ~200 Kb, which contrasts with the human genomic average of 1 gene per ~100 Kb) (IHGSC 2004). To

test whether the size of the L1 insertions at L1IMD loci showed any relation to its surrounding gene density, we performed correlation tests for each window size (4, 2 and 0.5 Mb) in both chimpanzee and human. While we found no significant correlation ($-0.16 < r < 0.34$, $P > 0.05$ in all cases), the *r*-value itself was negative in five out of six tests, opening the possibility that analysis of a larger dataset of L1 insertions may show a trend towards shorter L1 insertions in gene-rich areas of the genome. Because the chimpanzee LC23 locus was located in an unusually gene dense region in the short arm of chromosome 9 (***i.e.*** 1 gene per ~30 Kb), we repeated our correlation tests involving chimpanzee loci including and excluding this locus. However, the results were similar.

To characterize L1 insertions causing deletions within genes, we analyzed the 14 L1IMD loci (ten in human and four in chimpanzee) that were located within the introns of known or predicted genes. Eight of these were in collinear orientation with the gene transcript, while six were in antisense orientation. The average length of the L1 insertions within introns was considerably lower than the average L1 insertion length observed at non-intron L1IMD loci in both human and chimpanzee (849 vs. 1601 bp and 474 vs.1053 bp, respectively). These 47 % and 55% reductions, respectively, might indicate that smaller L1 insertions are better tolerated than longer ones within the introns of genes.

Discussion

The role of *Alu* and L1 retrotransposons in the creation of genomic instability is no longer a matter of dispute (Callinan et al. 2005; Gilbert et al. 2002; Kazazian and Goodier 2002; Symer et al. 2002). While extensive cell culture analyses have documented in detail the types and prevalence of genomic rearrangements by L1 insertion *in vitro*, the possibility remains that *in vivo*, evolutionary factors such as selection, variation in the number of actively retrotransposing

elements and differences in effective population size (Boissinot *et al.* 2001; Hedges *et al.* 2004) may substantially impact the spectrum of these rearrangements. To test the latter, we made use of the genome sequence of our closest living relative, the common chimpanzee (*Pan troglodytes*), and performed a human/chimpanzee comparison of L1IMD events.

Evolutionary Levels of L1IMD

The previous cell culture analyses of Symer et al. (2002) and Gilbert et al. (2002), have both reported the presence of large (> 3 Kb) deletions associated with L1 retrotransposition, with one candidate in Gilbert et al. (2002) even deleting at least 24 Kb and possibly as much as 71 Kb of target sequence. However, such massive deletions are very unlikely to persist in the population because of the likelihood that such events would delete regions of the genome required for survival and thus would subsequently be removed by selection. Consistent with this view, we find that the vast majority of L1IMDs with some degree of evolutionary success are shorter than a few hundred bases in both the human and chimpanzee lineages. In fact, the total amount of lineage specific deleted sequences through L1IMD in the latest draft of the human genome is estimated to be only ~17.7 Kb, corresponding to an average deletion rate of ~3.5 Kb per haploid genome per million years (Myrs) within the ~5 Myrs since the divergence of humans and chimpanzees (Chen and Li 2001; Goodman et al. 1998). The rate of deletion in the chimpanzee genome is also similar at ~3 Kb per haploid genome per Myrs.

To estimate the number of human-specific L1 insertions, we reasoned that all human-specific L1 elements belong to only 3 subfamilies (L1Ta, L1preTa and L1PA2) (Furano et al. 2004; Myers et al. 2002; Salem et al. 2003b). Given that both empirical (Boissinot *et al.* 2004) and theoretical (Hedges *et al.* 2004) evidence suggests that the analysis of a single genome results in the recovery of only ~50% of all polymorphic elements in a subfamily, we estimated

each L1 subfamily copy number as the sum of the number of fixed elements and twice the number of polymorphic elements detected in the human genome reference sequence. This resulted in a total of ~5800 L1 elements for these three subfamilies. However, not all of these L1 elements are specific to humans (Buzdin *et al.* 2003). Using the method of identification of human-specific L1 insertions from Buzdin et al. (2003), we conclude that ~1300 L1 elements have inserted in the human genome since the human/chimpanzee divergence. Given that L1 elements in the human genome have an average size of ~1 Kb (Lander et al. 2001), we calculate that the insertion of L1 elements within the past 5 Myrs resulted in the addition of ~1.3 Mb of sequence to the human genome. This is two orders of magnitude higher than the ~18 Kb length of sequence deleted in the same period by L1IMDs. On a larger time scale, assuming that ~2.2% of L1 insertions are associated with L1IMD in primates (29/1300 in humans) and the median deletion size of 21 bp from the L1IMD events in our study, the ~520,000 L1 elements that inserted in primate genomes were responsible for the deletion of a minimum of ~240 Kb of DNA sequences. However, if we perform the same calculation using the average L1IMD size of 655 bp, then almost 7.5 Mb of primate genomic DNA would have been deleted during the retrotransposition of L1 elements. It is also interesting to note that ~520 Mb (520,000 L1 elements with an average size of 1 Kb) of sequence has been added to the genome by the insertion of L1s in the same time period. This is reflective of the ongoing process of renewal of genomic sequences through the retrotransposition process.

Chronological Framework of L1IMD Events

We were able to place our L1IMD events in a chronological framework on the basis of (i) the results of the median-joining network analysis (Figure 2.3); (ii) the observation that about two thirds of the human-specific L1IMDs are caused by L1PA2 insertions vs. about one third

caused by members of the younger L1Hs subfamily; and (iii) only 20% of the chimpanzee-specific L1IMD events were specific to the common chimpanzee and 80% are shared with the pygmy chimpanzee. Taken together, these results suggest that L1IMD events in the human genome may have occurred to a large extent soon after the human/chimpanzee divergence when the L1PA2 subfamily was active, although they may be continuing to accumulate, as suggested by the non-trivial contribution of L1Hs members. In the chimpanzee lineage as well, the majority of L1IMDs is older than 1-2 Myrs, which corresponds to the divergence time of common and pygmy chimpanzees (Chen and Li 2001; Goodman et al. 1998). However, these observations may, at least partly, be influenced by the overrepresentation of older insertions within genomic sequences (i.e. younger events are more likely to be polymorphic than older events and could remain undetected when a small number of individuals were sequenced). Nevertheless, the fact that 23 out of 26 L1IMDs in the common chimpanzee involve L1PA2 elements suggests that the L1PA2 subfamily may still be actively undergoing retrotransposition in the chimpanzee lineage.

Interestingly, among the chimpanzee-specific L1IMDs, we found an ancient L1PA5 element (LC8) that was polymorphic. The L1PA5 subfamily is ~25 Myrs old (Furano *et al.* 2004). We excluded the possibility of polymorphism being maintained by balancing selection acting on this locus because of the low gene density in its vicinity. It is worthy to note that Bennett et al. (2004) also recently identified four polymorphic old *Alu*S elements and one L1PA3 polymorphism. Therefore, this suggests that at least some copies of older L1 retrotransposon subfamilies can retain the ability of retrotransposition for extended periods of time similar to *Alu* elements (Han et al. 2005). Alternatively, it is possible that these polymorphisms have been maintained over a very long period of time by chance. Although this is expected to happen very rarely, it may not be surprising to find a few such cases in view of the hundreds of thousands of L1 and *Alu* elements (Batzer and Deininger 2002; Britten 1994; Furano et al. 2004) that have

inserted during primate evolution. However, we favor the former explanation in the case of the polymorphic L1PA5 element we detected, because DNA sequencing of the locus showed that the L1PA5 insert was specific to the chimpanzee lineage and absent from all other primate genomes we examined.

Different Mechanisms May Exist for Different Deletion Sizes

The sizes of the L1IMDs we identified are in general agreement with the size range of similar deletions (13 deletion events ranging from 2 bp-14 Kb) identified in a recent study of L1 retrotransposition in cell culture (Gilbert *et al.* 2005). However, our sample size for L1IMDs is substantially larger. Very large deletions like those seen in cell culture analyses (Gilbert *et al.* 2002; Symer *et al.* 2002) did not appear in our study, presumably because they are more likely to have been removed from the populations rapidly due to their deleterious nature (especially if they were located in gene-rich regions). Interestingly, in both the human and chimpanzee datasets, we noticed a tendency for the deletions to be either very short (*i.e.* < 100 bp) or, to a lesser extent, relatively large (> 1 Kb), which possibly indicates the concomitant action of two different mechanisms of L1IMD acting on different scales. This dichotomy in deletion sizes was also observed by Gilbert et al. (2002), and our data would seem to fit their general models for small and large L1IMD events, to which we propose further extensions to better explain some of the L1 structures that are unique to our study. In general, small deletions may be caused by the creation of 5' overhangs by top strand cleavage being inexactly opposed to bottom strand cleavage in an upstream direction, with subsequent 5'-3' exonuclease activity on both the exposed 5' ends (Figure 2.5A). By contrast, larger deletions may be explained if the nascent L1 cDNA invades a double-strand break with a 3' overhang located upstream to the initial integration site (Figure 2.5B), with gap repair removing the intervening single-stranded segment

and causing a large deletion (Gilbert *et al.* 2002; Gilbert *et al.* 2005). Additionally, we suggest that large deletions could result if palindromic stretches downstream of the original site of integration, mechanically or enzymatically held in single-strand conformation during the physical integration of the L1 DNA, formed hairpin loops which were subsequently removed by repair enzymes. Remarkably, a similar pattern of deletion size differences (small or large) also characterizes the deletions caused in the target sequence by the retrotransposition of *Alu* elements (Callinan *et al.* 2005). Taken together, the data from genomic deletions caused by L1 and *Alu* retrotransposon insertions are consistent with the view that two different mechanisms underlie the deletions of small and large stretches of target sequence, especially as both *Alu* retrotransposition-mediated deletions (Callinan *et al.* 2005) and the L1IMDs in our study are whole-genome analyses that should represent the comprehensive picture of such deletions. **A Model for Correlation between Insert Size and Deletion Size**

In both our human and chimpanzee data sets, we noted a significant positive correlation between the size of the L1 insertion and the size of the deletion caused thereupon. In the extension of the model of Gilbert et al. (2002) described above for the creation of large deletions, we propose a probability-based mechanism to further explain the observed correlation (Figure 2.5B). Our model assumes that given the prior presence of a 3' overhang in the double-strand break (which is a necessary prerequisite for the occurrence of the deletion by this mechanism) a longer segment of newly transcribed minus strand L1 cDNA is more likely to contain the adequate number of complementary bases (and thus be able to bind with sufficient strength) than a shorter segment. A longer stretch of complementarity than expected by chance between the end of the L1 cDNA and the region surrounding the 5' end of the L1 insertion in the ancestral (pre-insertion) sequence would provide support for this model. To quantify this parameter, we located

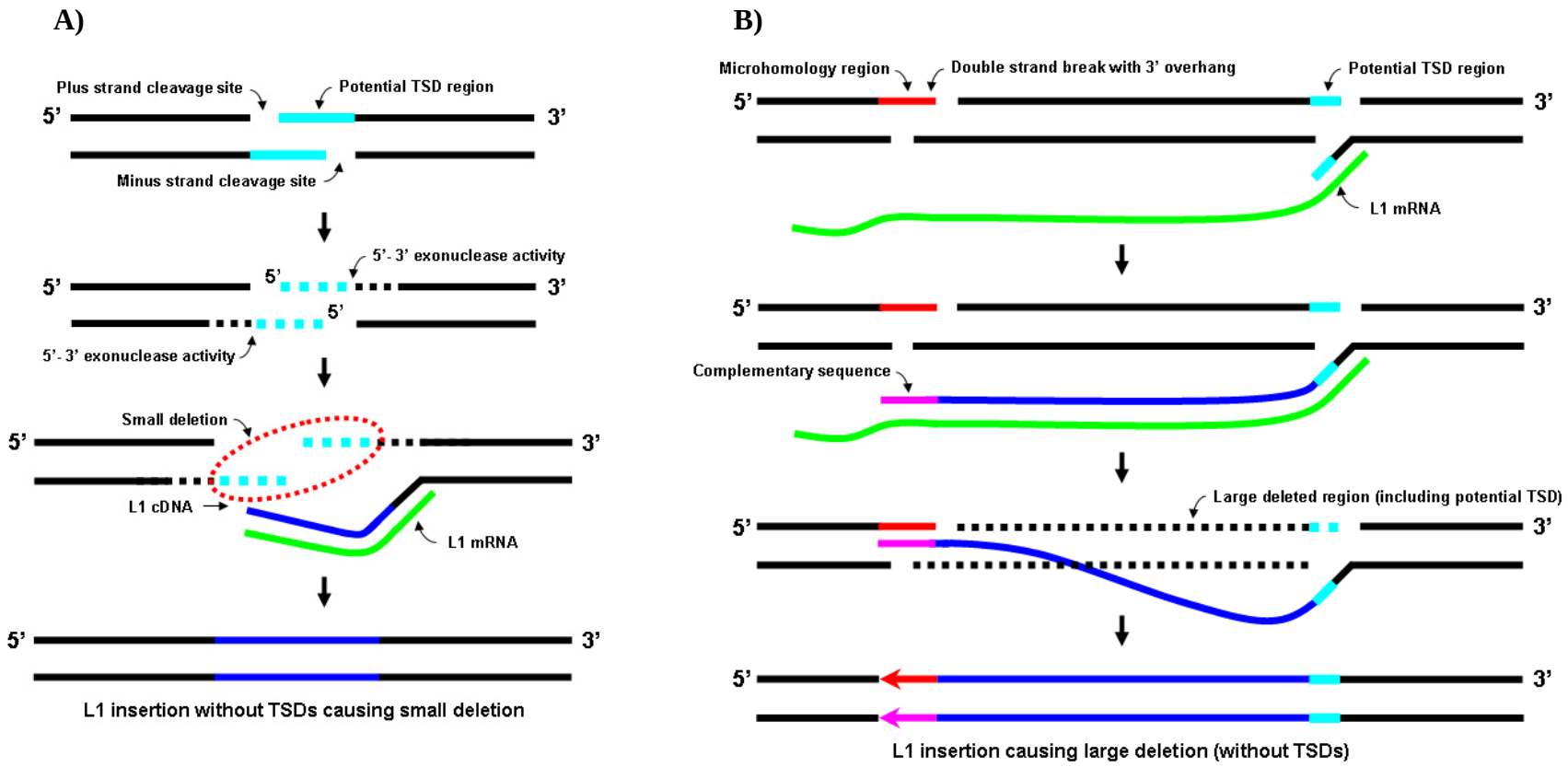


Figure 2.5. Models for the creation of L1IMDs. (A) Formation of small deletions. 5' overhangs created by inexact cleavage of the top strand by the L1 EN are subject to 5'-3' exonuclease activity, that removes small single-stranded stretches from both the plus and minus strands (dotted light blue lines), which would otherwise have been the templates for the formation of TSDs. Subsequent ligation of the L1 cDNA to the upstream minus strand sequence and plus strand sequence synthesis by cellular enzymes results in the creation of small deletions and an L1 insertion without TSDs. (B) Formation of large deletions. For any preexisting double strand break that has a 3' overhang (red) for base pairing of the L1 cDNA (blue), a longer cDNA transcript is more likely to contain a stretch of sequence that has adequate complementary bases for annealing (pink) than a shorter one. Subsequent recombinational repair would remove a large segment of the target sequence, extending downstream to the original integration site (dotted black line) and resulting in a L1 insertion without TSDs.

(a) the 5' start position of the L1 insertions with respect to the L1.3 consensus sequences and; (b) the site corresponding to the 5' start position of the human-specific L1 insertions in the chimpanzee genomic sequence and *vice versa.* Next, we isolated 15 bp stretches of sequence in the 5' direction from both these locations in the L1.3 consensus sequence and the genomic sequences, respectively, and aligned them. In all the 12 L1IMD loci that had large deletions corresponding to large L1 insertions (both sizes above 500 bp), we found between 27% and 53% complementary bases, which would indicate that potential binding sites were present in all the cases (see supplementary data for alignments in Batzer Laboratory Web site). Additionally, in seven out of the 15 loci, the first two (LH28, LH30, LC4, LC31) to three (LH17, LH27, LC29) bases in the 3' end of the alignments were complementary. This further indicates that these bases could have been utilized for binding between the L1 transcript and the target sequence. Recent computational analyses of the 5' junctions of young L1 insertions in the human genome (Zingler et al. 2005) suggest that microhomology-mediated end-joining is the likely mechanism for 5'-end attachment during the retrotransposition of 5'truncated L1 elements. Thus, our results support this hypothesis and indicate that longer L1 cDNA strands, because of the higher probability of possessing such microhomology with the pre-integration site, are better suited to the creation of longer genomic deletions by bridging double strand breaks. The presence of two double strand breaks (one at the original integration site and one upstream of it) would also lessen the chance of mechanical obstruction to the annealing of the L1 cDNA across the potential deleted region. We note that as proposed in Gilbert et al. (2005), the site of integration is very likely to be a "host/parasite battleground", with the L1 cDNA trying to finish reverse transcription and the host enzymatic machinery opposing it. Given the odds against the simultaneous occurrence of L1 insertion reaching comparatively near full-length and the

presence of a double-strand break with a 3' overhang conducive to binding, the lower number of large deletions corresponding to large insertions (6/26 in chimpanzee and 6/24 in human) lends support to our model.

Rearrangements within the L1 Elements Associated with L1IMD.

Six of the L1IMD loci were also characterized by rearrangements within the sequence of the L1 insertion, resulting in atypical L1 structures. Of these, two were both 5' truncated and partially inverted (LC26 and LC27) while the other four (LH17, LH19, LH26 and LH31) were 5' truncated non-inverted L1 elements that showed internal rearrangements. Previous cell culture studies have also shown that L1 rearrangements can occur during the process of retrotransposition (Gilbert *et al.* 2002; Gilbert *et al.* 2005). In our study, the presence of the homologous sequence from the respective closest ancestors allowed us to confirm that these loci did not have prior insertions of endogenous L1 elements at the pre-integration sites. The probability of two independent L1 insertions into the same locus after the human-chimpanzee divergence is extremely small, given the large size of the human and chimpanzee genomes and the estimated number of L1 insertions specific to these lineages (e.g. ~1300 in humans), which leads us to suggest that mechanistic processes led to the generation of these particular structures during the retrotransposition events. Of the non-inverted atypical L1 elements, LH19 and LH26 are strong candidates for gene duplication, with portions of the L1.3 consensus sequence repeated in parallel orientation without any intervening region (53 and 189 bp, respectively). LH17 and LH31 were 5' truncated L1 insertions that showed two stretches of the consensus L1.3 sequence with a gap of ~300 bp in between them. We propose a novel mechanism for this structure, by which stretches of microhomology within the L1.3 consensus sequence might have led to the L1 mRNA looping back on itself (Figure 2.6A), resulting in the formation of an L1

insertion with the characteristic structure observed and an associated deletion of target site DNA. The presence of at least one such 8 bp homologous stretch was visually confirmed by us in both the cases.

With respect to the 5' truncation/inversions in our study (LC26 and LC27), a mechanism termed 'twin priming' has been suggested for the creation of such structures during L1 retrotransposition (Ostertag and Kazazian 2001b). However, the existing model does not incorporate the possibility of creation of large deletions during this process. To provide a possible explanation for the large deletions caused at these loci (2973 and 1175 bp, respectively), we suggest a 'modified twin priming' model, whereby a stretch of complementarity between the extended L1 mRNA and a 3' overhang formed at a preexisting double strand break would lead to a second site of priming on the mRNA (Figure 2.6B). Subsequently, dissociation of the two newly synthesized cDNA segments from the mRNA and the formation of an 'inversion junction', followed by double strand synthesis, would lead to the removal of the intervening DNA (between the original site of TPRT and the double strand break) with the formation of a rearranged L1 element with the truncation/inversion structure observed.

Conclusion

In conclusion, our study demonstrates that L1IMDs are not restricted to transformed cells but are also a feature of *in vivo* insertions as well, and that this process has been active in causing deletions in both the human and chimpanzee lineages. Our *in vivo* evolutionary analysis and prior *in vitro* cell culture studies of deletions caused by L1 retrotransposition provide pictures that differ at first sight, but can be reconciled by evolutionary factors. While 16-25% of L1 insertions identified in the cell culture studies cause deletions at the target site (Gilbert et al. 2002; Gilbert et al. 2005; Symer et al. 2002), only ~2.2% of existing human-specific L1

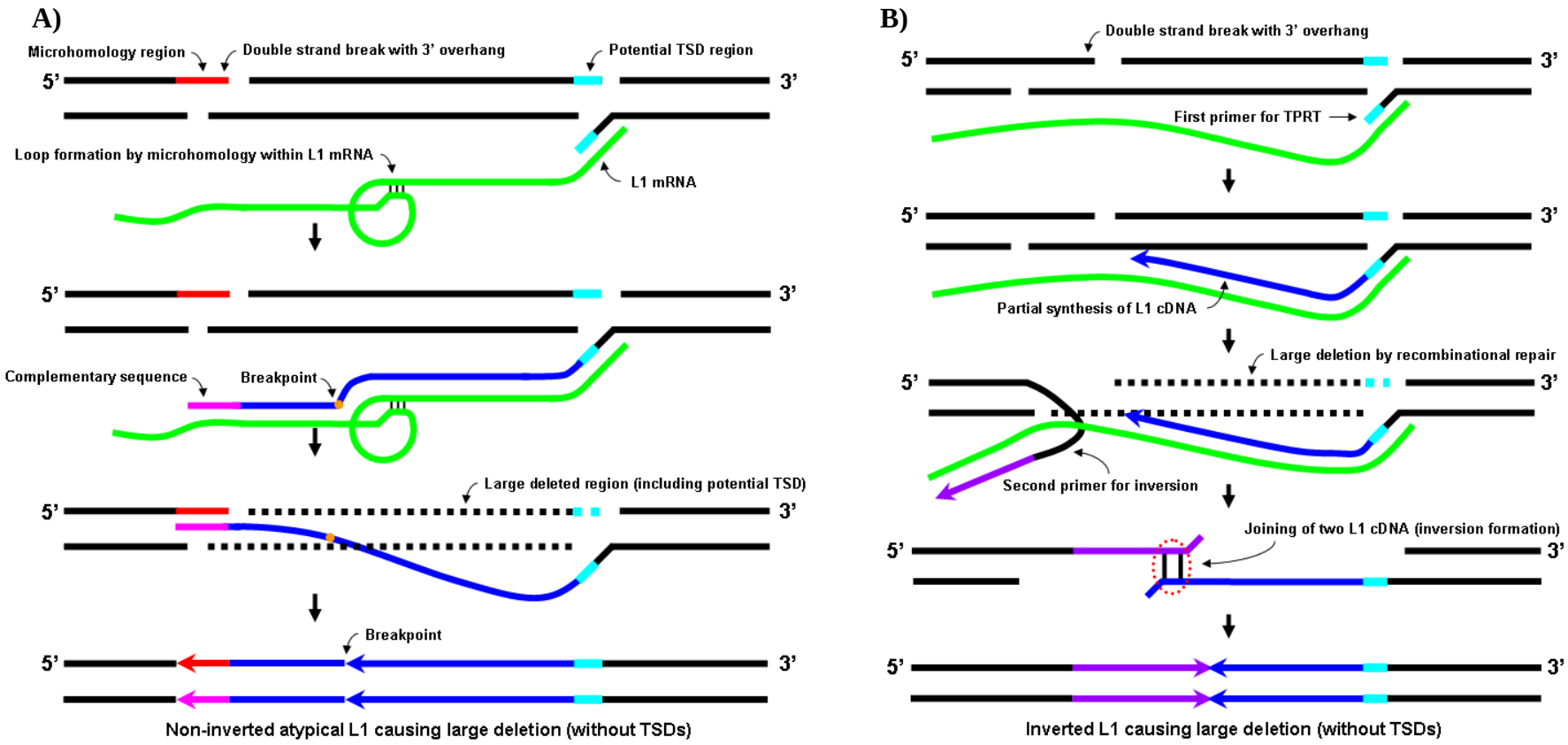


Figure 2.6. Models for the formation of deletions associated with atypical L1 elements. (A) Formation of a non-inverted atypical L1 insertion resulting in a large deletion. The L1 mRNA (green) forms a loop, with microhomology stretches within its sequence annealing to each other. The resulting L1 cDNA (blue) has an internal breakpoint (orange) where a stretch of the consensus sequence (complementary to the loop) is missing. Arrows show the orientation of the two parts of the L1 insertion. (B) Formation of a 5' truncation/inversion resulting in a large deletion. Annealing of the L1 mRNA (green) to a complementary sequence in the 3' overhang of a preexisting double-strand break leads to the transcription of a second stretch (purple) apart from the original cDNA (blue). Subsequently, both dissociate from the mRNA and form an 'inversion junction' (circled in red). Recombinational repair removes the stretch of DNA between the double strand break and the original site of integration. Plus-strand synthesis results in a 5' truncated L1 with the inverted portion being reverse complementary to the consensus sequence. Arrows show the orientation of the L1 segments in the inversion.

insertions seem to be directly linked to genomic deletions [compared to 0.2-0.4% for *Alu* elements (Callinan *et al.* 2005)]. As the currently available chimpanzee assembly covers ~95% of the genome sequence while the human genome sequence is considered to be "finished" (UCSC genome database), our human-chimpanzee comparison probably recovered most species-specific L1IMD events. A slight underestimation due to different levels of completion of the human and chimpanzee genome sequences could not account for the ~10-fold difference between *in vivo* and *in vitro* L1IMD rates. The difference in the rate of L1IMD estimated from cell culture-based analyses and genome-based analyses may more likely reflect the differences in the number of these events that are tolerated in the genome after natural selection has occurred. Thus, our study validates the use of cell culture retrotransposition assays as surrogate models to deduce the underlying mechanisms for these complex genomic rearrangements.

The extent of genomic deletion is reduced compared to the amount of sequence inserted by the L1 retrotransposition process. In addition evidence from our study indicates that many large L1IMDs such as those identified in cell culture assays do not persist in the primate lineage over time. We propose new mechanisms for the creation of some of the specific L1 structures reported in our analysis. Most of the existing human-specific deletions appear to have taken place soon after the divergence of the human and chimpanzee lineages. The atypical L1 elements created during the deletion process could also be sources for new L1 subfamilies in both the human and chimpanzee lineages (Gilbert et al. 2002; Saxton and Martin 1998).

Materials and Methods

Computational Analysis

To identify L1IMD candidate loci in the human genome, we first identified all L1 elements that have intact 3' sequence in the July 2003 freeze of the human genome (hg16: UCSC

genome database at http://genome.ucsc.edu/ENCODE/) by querying the genome sequence with the 50 bp of the 3'-end of the L1 consensus sequence (excluding the poly(A) tail), using the command line version of the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.* 1990). The BLAST output file was then processed by a set of in-house Perl programs to extract entries that contain matches with at least 96% sequence similarity to the query sequence over at least 40 bp, resulting in a total of 49,791 L1 entries. Using a cutoff value of 96% similarity ensured that the most recent L1 inserts (including human-specific events) were selected for further analysis. For each entry, 400 bp of sequence downstream of the start of the query (including the match to the query sequence, the poly(A) tail and the 3'-end flanking sequence) were extracted from the human genome sequence. The exact start of the 3'-end flanking sequences was determined for each entry by aligning it with the 50 bp L1 consensus sequence used as the initial query, with which a stretch of 100 adenosines was now included to simulate the poly(A) tail. The 3' sequence immediately flanking the L1 element identified for each entry was then used as a query to search the chimpanzee genome (PanTro1; Nov. 2003 freeze). If the best match started immediately after the poly(A) tail, the locus was considered to be a human-specific L1 insertion and the start of the matching region was considered to be the insertion site in the human genome. For each identified locus, we extracted 1000 bp and 100 bp of sequence in the 5' and 3' regions of the pre-insertion site, respectively, from the chimpanzee genome. The 5' chimpanzee sequences were then used to query the human genome. If a 1000 bp chimpanzee sequence only matched the human sequence at its 5' end, the unmatched sequence at the 3' end was considered as a L1IMD candidate in the human genome. In cases where there was no match in the entire 1000 bp of the query sequence, the 5' flanking sequences from the chimpanzee genome were progressively extended until a good partial match at the 5' end could be identified

in the human sequence. These cases were considered to represent deletions that were close to or longer than 1000 bp.

Chimpanzee L1IMD candidates were identified by reversing the query and target genomes and using the same approach as described above. All candidate loci were then subjected to manual verification, resulting in a total of 30 and 33 putative L1IMDs in the human and chimpanzee genomes, respectively.

PCR Amplification and DNA Sequence Analysis

To experimentally verify the L1IMD candidate loci, flanking oligonucleotide primers were designed using the primer design software Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The primers were subsequently screened against the GenBank NR and HTGS databases using BLAST queries to determine if they resided in unique DNA sequences. Detailed information for each locus including primer sequences, annealing temperature, PCR product sizes and chromosomal locations can be found in the "Publications" section of our website (http://batzerlab.lsu.edu).

PCR amplification of each locus was performed in 25 µl reactions using 10-50 ng DNA, 200 nM of each oligonucleotide primer, 200 µM dNTP's in 50 mM KCl, 1.5 mM MgCl2, 10 mM Tris-HCl (pH 8.4) and 2.5 units *Taq* DNA polymerase. Reactions were subjected to an initial denaturation step of 94° C for four minutes, followed by 32 cycles of one minute of denaturation at 94° C, one minute of annealing at optimal annealing temperature and one minute of extension at 72° C, followed by a final extension step at 72° C for ten minutes on a Biorad™ iCycler thermocycler. Resulting PCR products were separated on 2% agarose gels, stained with ethidium bromide and visualized using UV fluorescence.

Individual PCR products were purified from the gels using the Wizard® gel purification kit (Promega) and cloned into vectors using the TOPO-TA Cloning® kit (Invitrogen). For each sample, three colonies were randomly selected and sequenced on an Applied Biosystems AB3100 automated DNA sequencer using chain termination sequencing (Sanger *et al.* 1977). All clones were sequenced in both directions using M13 forward and reverse primers to confirm the sequence, analyzed using the Seqman™ program in the DNASTAR suite and aligned using the BioEdit sequence alignment software package (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

For each locus, this procedure was applied to one individual from each of five different primate species, including *Homo sapiens* (HeLa cell line ATCC CCL-2), *Pan troglodytes* (common chimpanzee; cell line AG06939B), *Pan paniscus* (bonobo or pygmy chimpanzee; cell line AG05253B), *Gorilla gorilla* (Western lowland gorilla; cell line AG05251) and *Pongo pygmaeus* (orangutan; cell line ATCC CR6301). The DNA sequences from this study are available in GenBank under accession numbers DQ017967-DQ018078.

Polymorphism Analysis

To evaluate the extent of polymorphism associated with the validated L1IMD loci, each locus was further amplified in the genomes of 80 humans (20 individuals from each of four populations, see below) and 12 unrelated common chimpanzees, following the PCR protocol described above. Our human population panel was composed of DNA from African-American, European and Asian populations (isolated from peripheral blood lymphocytes) available from previous studies in our lab and South American population DNA (HD17 and HD18) purchased from the Coriell Institute for Medical Research. The common chimpanzee population panel was prepared from genomic DNA of twelve unrelated individuals of unknown geographic origin and

subspecies affiliation, which was provided by the Southwest Foundation for Biomedical Research.

Phylogenetic Analysis of L1IMDs

To examine the phylogenetic relationships of the human and chimpanzee L1 elements identified in this study, we constructed a median-joining network (Cordaux et al. 2004; Han et al. 2005) using the software NETWORK ver. 4.1.1.0 (Bandelt *et al.* 1999) available at http://www.fluxus-engineering.com/sharenet.htm. The network was generated using a 94 bp stretch corresponding to positions 5930-6023 in the 3' end consensus sequence of the L1Hs and L1PA2 reference sequences obtained from the RepeatMasker database. Elements LC9 and LH29 had to be excluded from this analysis because of truncations in the region analyzed.

Analysis of Flanking Sequences

For GC content analysis, we used the BLAST-Like Alignment Tool (BLAT) server (Kent 2002) available at http://genome.ucsc.edu/cgi-bin/hgBlat to isolate 20 Kb of flanking sequence in either direction from the reference human and chimpanzee draft sequences after adjustment at the 3' end to prevent bias towards excessive adenosine residues (see results). We used the EMBOSS GeeCee server (http://emboss.sourceforge.net/apps/geecee.html) to calculate GC percentages. To characterize the gene-frequency neighborhoods of the L1IMDs, we pinpointed exact chromosomal location of the L1 insertions with BLAT, and then used the NCBI MapViewer interface (http://www.ncbi.nlm.nih.gov/mapview/) to map all known genes within 4, 2 and 0.5 Mb windows surrounding the 5' and 3' ends of the L1IMDs.

References

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Bandelt, H.J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16:** 37-48.

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Boeke, J.D. and S.E. Devine. 1998. Yeast retrotransposons: finding a nice quiet neighborhood. *Cell* **93:** 1087-1089.

Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18:** 926-935.

Boissinot, S., A. Entezam, L. Young, P.J. Munson, and A.V. Furano. 2004. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res* **14:** 1221-1231.

Britten, R.J. 1994. Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. *Proc Natl Acad Sci U S A* **91:** 6148-6150.

Brouha, B., J. Schustak, R.M. Badge, S. Lutz-Prigge, A.H. Farley, J.V. Moran, and H.H. Kazazian, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100:** 5280-5285.

Burwinkel, B. and M.W. Kilimann. 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277:** 513-517.

Buzdin, A., S. Ustyugova, E. Gogvadze, Y. Lebedev, G. Hunsmann, and E. Sverdlov. 2003. Genome-wide targeted search for human specific and polymorphic L1 integrations. *Hum Genet* **112:** 527-533.

Callinan, P.A., J. Wang, S.W. Herke, R.K. Garber, P. Liang, and M.A. Batzer. 2005. Alu Retrotransposition-mediated Deletion. *J Mol Biol* **348:** 791-800.

Chen, F.C. and W.H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444-456.

Cordaux, R., D.J. Hedges, and M.A. Batzer. 2004. Retrotransposition of Alu elements: how many sources? *Trends Genet* **20:** 464-467.

Cost, G.J. and J.D. Boeke. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37:** 18081-18093.

Cost, G.J., A. Golding, M.S. Schlissel, and J.D. Boeke. 2001. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res* **29:** 573-577.

Dewannieux, M., C. Esnault, and T. Heidmann. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35:** 41-48.

Dombroski, B.A., A.F. Scott, and H.H. Kazazian, Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc Natl Acad Sci U S A* **90:** 6513-6517.

Feng, Q., J.V. Moran, H.H. Kazazian, Jr., and J.D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905-916.

Furano, A.V., D.D. Duvernell, and S. Boissinot. 2004. L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet* **20:** 9-14.

Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Gilbert, N., S. Lutz, T.A. Morrish, and J.V. Moran. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25:** 7780-7795.

Goodier, J.L., E.M. Ostertag, and H.H. Kazazian, Jr. 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* **9:** 653-657.

Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C.P. Groves. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9:** 585-598.

Han, K., J. Xing, H. Wang, D.J. Hedges, R.K. Garber, R. Cordaux, and M.A. Batzer. 2005. Under the genomic radar: The Stealth model of Alu amplification. *Genome Res* **15:** 655-664.

Hedges, D.J., P.A. Callinan, R. Cordaux, J. Xing, E. Barnes, and M.A. Batzer. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* **14:** 1068-1075.

Ho, H.J., D.A. Ray, A.H. Salem, J.S. Myers, and M.A. Batzer. 2005. Straightening out the LINEs: LINE-1 orthologous loci. *Genomics* **85:** 201-207.

IHGSC. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94:** 1872-1877.

Kazazian, H.H., Jr. and J.L. Goodier. 2002. LINE drive. retrotransposition and genome instability. *Cell* **110:** 277-280.

Kazazian, H.H., Jr. and J.V. Moran. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet* **19:** 19-24.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Luan, D.D., M.H. Korman, J.L. Jakubczak, and T.H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72:** 595-605.

Mathias, S.L., A.F. Scott, H.H. Kazazian, Jr., J.D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808-1810.

Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530-1534.

Morrish, T.A., N. Gilbert, J.S. Myers, B.J. Vincent, T.D. Stamato, G.E. Taccioli, M.A. Batzer, and J.V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31:** 159-165.

Myers, J.S., B.J. Vincent, H. Udall, W.S. Watkins, T.A. Morrish, G.E. Kilroy, G.D. Swergold, J. Henke, L. Henke, J.V. Moran et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71:** 312-326.

Nachman, M.W. and S.L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156:** 297-304.

Ostertag, E.M. and H.H. Kazazian, Jr. 2001a. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35:** 501-538.

Ostertag, E.M. and H.H. Kazazian, Jr. 2001b. Twin priming: a proposed mechanism for the creation of inversions in l1 retrotransposition. *Genome Res* **11:** 2059-2065.

Ovchinnikov, I., A. Rubin, and G.D. Swergold. 2002. Tracing the LINEs of human evolution. *Proc Natl Acad Sci U S A* **99:** 10522-10527.

Ovchinnikov, I., A.B. Troxel, and G.D. Swergold. 2001. Genomic Characterization of Recent Human LINE-1 Insertions: Evidence Supporting Random Insertion. *Genome Res* **11:** 2050-2058.

Pickeral, O.K., W. Makalowski, M.S. Boguski, and J.D. Boeke. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* **10:** 411-415.

Roy-Engel, A.M., A.H. Salem, O.O. Oyeniran, L. Deininger, D.J. Hedges, G.E. Kilroy, M.A. Batzer, and P.L. Deininger. 2002. Active Alu element "A-tails": size does matter. *Genome Res* **12:** 1333-1344.

Salem, A.H., G.E. Kilroy, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003a. Recently integrated Alu elements and human genomic diversity. *Mol Biol Evol* **20:** 1349-1361.

Salem, A.H., J.S. Myers, A.C. Otieno, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003b. LINE-1 preTa elements in the human genome. *J Mol Biol* **326:** 1127-1146.

Salem, A.H., D.A. Ray, and M.A. Batzer. 2005. Identity by descent and DNA sequence variation of human SINE and LINE elements. *Cytogenet Genome Res* **108:** 63-72.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Sassaman, D.M., B.A. Dombroski, J.V. Moran, M.L. Kimberland, T.P. Naas, R.J. DeBerardinis, A. Gabriel, G.D. Swergold, and H.H. Kazazian, Jr. 1997. Many human L1 elements are capable of retrotransposition. *Nat Genet* **16:** 37-43.

Saxton, J.A. and S.L. Martin. 1998. Recombination between subtypes creates a mosaic lineage of LINE-1 that is expressed and actively retrotransposing in the mouse genome. *J Mol Biol* **280:** 611-622.

Smit, A.F., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246:** 401-417.

Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Szak, S.T., O.K. Pickeral, W. Makalowski, M.S. Boguski, D. Landsman, and J.D. Boeke. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052.

Watanabe, H., A. Fujiyama, M. Hattori, T.D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382-388.

Wei, W., N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian, J.D. Boeke, and J.V. Moran. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21:** 1429-1439.

Zingler, N., U. Willhoeft, H.-P. Brose, V. Schoder, T. Jahns, K.-M.O. Hanschmann, T.A. Morrish, J. Löwer, and G.G. Schumann. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15:** 780-789.

CHAPTER THREE:

HUMAN GENOMIC DELETIONS MEDIATED BY RECOMBINATION BETWEEN *ALU* ELEMENTS*

*Reprinted by permission of American Journal of Human Genetics

Introduction

With a copy number of >1 million, *Alu* elements are one of the most successful non-LTR retrotransposon families in the human genome (Lander et al. 2001a). In addition to classic retrotransposition-associated insertion mutations, *Alu* elements can create genomic instability by the deletion of host DNA sequences during their integration into the genome and by creating genomic deletions associated with intrachromosomal and interchromosomal recombination events (Callinan et al. 2005; Deininger and Batzer 1999). Multiple features predispose *Alu* elements to successful recombination, including their proximity in the genome (one insertion every 3 kb on average), the high GC content of their sequence (~62.7%), and the remarkable sequence similarity (70-100%) among *Alu* subfamilies of widely different ages. Overall, the recombinogenic nature of these elements is reflected in the various forms of cancer and genetic disorders associated with *Alu*-mediated recombination events (Batzer and Deininger 2002; Deininger and Batzer 1999; Hattori et al. 1999; Huang et al. 1989; Levran et al. 1998; Marshall et al. 1996; Myerowitz and Hogikyan 1987; Rohlfs et al. 2000; Rothberg et al. 1997; Tvrdik et al. 1998).

However, clinical studies of isolated disease-causing deletions, although useful from a medical viewpoint and in demonstrating the existence of *Alu* Recombination-Mediated Deletions (ARMDs), do not adequately depict the overall contribution of this process to the architecture of the genome and the associated impact on gene function. The availability of a genome sequence for the common chimpanzee (*Pan troglodytes*), the closest evolutionary relative of the human lineage (CSAC 2005), has allowed us to perform a comparative genomic assessment of the extent of ARMD in the human genome over the past ~6 million years, since the divergence of the human and chimpanzee lineages (Miyamoto *et al.* 1987; Wildman *et al.* 2003). In this study,

we identified ~400 kb of human-specific ARMD, the distribution of which is biased toward gene-dense regions of the genome, which raises the possibility that ARMD may have played a role in the divergence of humans and chimpanzees. About 60% of the ARMDs are located in genes, and, in at least three instances, exons have been deleted in human genes relative to their chimpanzee orthologs. The nature of the altered genes suggests that ARMD might have played a role in shaping the unique traits of the human and chimpanzee lineages. Mechanistically, we characterized the physical aspects of the deletion process and proposed different models for ARMD.

Results

A Whole-Genome Analysis of Human-Specific ARMD Events

To identify putative ARMD loci, we first computationally compared the human and chimpanzee genomes. Subsequently, we manually inspected and, if needed, experimentally verified individual loci. Of the 1332 computationally predicted deletions that we initially recovered, 461 were discarded after manual inspection (Table 3.1). The causes for rejection of computationally predicted ARMD loci were: (a) insertion of an *Alu* or other retroelement at the

Table 3.1. Summary of human-specific ARMD events

Classification	No. of loci
Computationally predicted deletion loci	1322
Discarded after manual inspection	461
Candidate ARMD events:	871
False-positive events (*Alu* insertion in chimpanzee):	379
Confirmed by PCR analysis	189
Analysis based on TSD structure	190
ARMDs:	492
Confirmed by PCR analysis	163
Analysis based on TSD structure	329

orthologous chimpanzee locus, which leads to the presence of sequence that the computer erroneously assumed to be deleted in the human genome (38 cases), (b) authentic deletion products in the human genome that were not products of *Alu-Alu* recombination (211 cases), and (c) computational errors in alignment of the human and chimpanzee genomes (212 cases). On the basis of sequence architecture, the remaining 871 loci represented putative ARMD events in the human lineage. All of these loci were further manually inspected and were analyzed, for comparison of the ancestral predeletion and human postdeletion states, by use of a TSD-based strategy as described below (see Materials and Methods). In addition, we experimentally verified the authenticity of 352 candidate ARMD loci by PCR (Table 3.1 and Figure 3.1). To be conservative, we discarded all loci in which an alternative mechanism (*e.g.*, random genomic

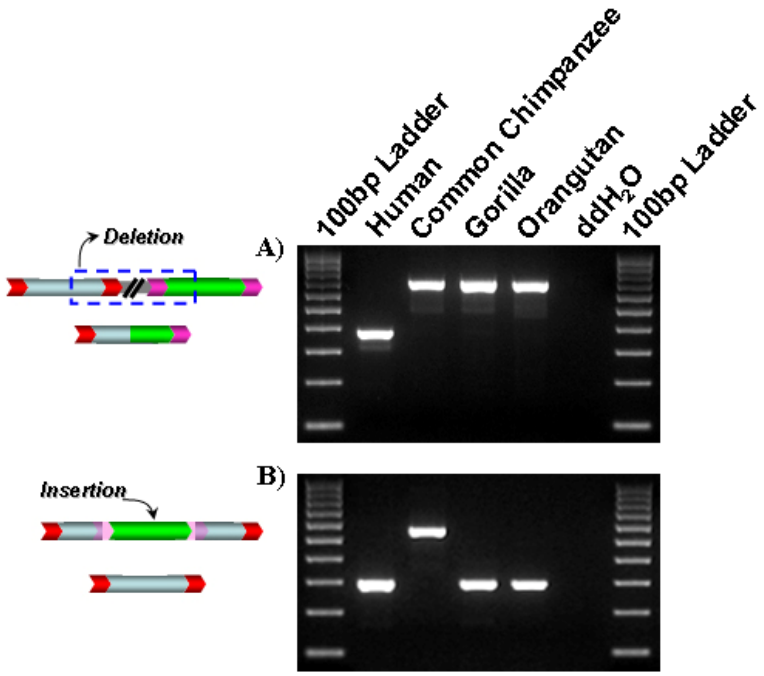


Figure 3.1. ARMD in the human genome. Individual ARMD candidate loci amplified by PCR. (A) Agarose-gel chromatograph of PCR products derived from an authentic human-specific ARMD event. (B) Agarose-gel chromatograph of PCR products derived from an ARMD false positive event (*Alu* insertion in chimpanzee). The DNA templates used in each reaction are shown above the chromatographs.

deletion), distinct from ARMD, could have produced the deletion. Specifically, ARMD events can be distinguished from random genomic deletions occurring at *Alu* insertion sites because an ARMD event reconstitutes an uninterrupted chimeric *Alu* element *(i.e.,* with no internal deletion), whereas the probability of this happening through chance alone (as would be the case with a random deletion) is remote. Indeed, the probability of two ~280-bp *Alu* elements breaking by chance at a homologous site is only 1 in ~80,000 (1 in 280 × 1 in 280). Hence, although we cannot formally exclude the possibility that a few random deletions may precisely mimic the ARMD process, we believe the overall impact of these nonauthentic events on our estimates would be minimal.

The manual verification of the 871 loci resulted in a final dataset of 492 ARMD events spanning the entire human genome (Table 3.1). Nine ARMD loci on the Y chromosome were all located in the pseudoautosomal part of this chromosome and hence were identical copies of deletion loci on the X chromosome. As a result, each event was counted only once during the analysis. In general, the loci analyzed in this study suggest that the combination of computational data mining and experimental validation is the "gold standard" when conducting comparative genomic searches for lineage-specific deletions. As we observed during the course of this study, lineage-specific insertions in one genome stand a risk of being characterized as deletions in the other when only two genomes are compared in a computational analysis. In our analysis, we minimized the chances of including such events by using three other hominoid genomes as controls during experimental verification of the events.

Extent of Genomic Deletion and Size Distribution of ARMD Events

The number of ARMD events is positively correlated with the number of *Alu* elements present on each chromosome ($r = 0.69$; $P < .0005$). This is expected, since physical proximity

between repetitive elements strongly predisposes them to recombination (Inoue and Lupski 2002). Simultaneous mapping of ARMD loci and all *Alu* insertions on each chromosome highlights the tendency for deletions to cluster with regions of high local *Alu* density (Figure 3.2). Additionally, sequence analysis of the *Alu* elements involved in ARMD events indicates that the number of elements from each *Alu* subfamily (Figure 3.3) is proportional to their genomewide copy number (Batzer and Deininger 2002), with no bias observed for elements from

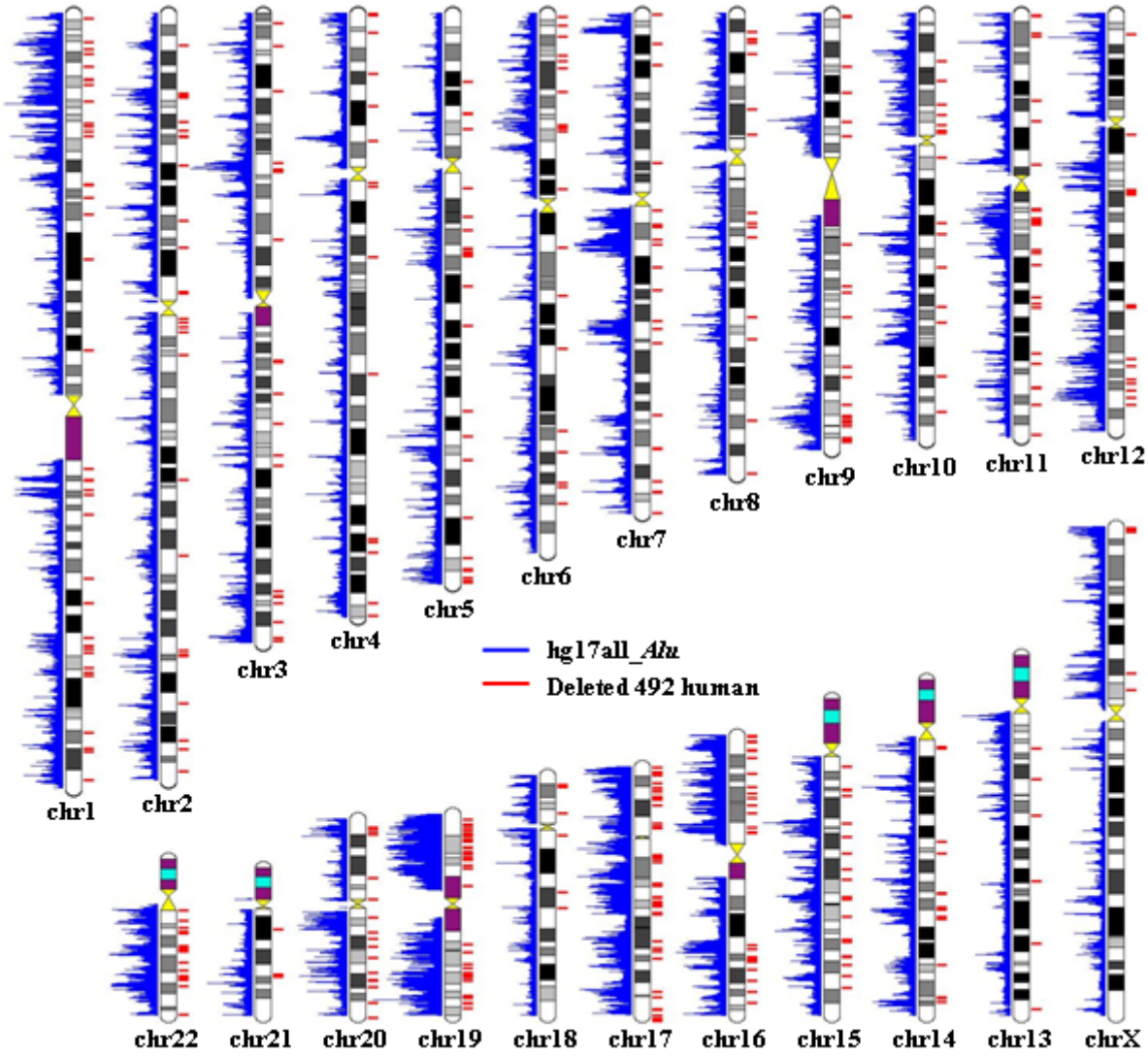


Figure 3.2. Density of ARMD events (red lines) and all *Alu* insertions (blue lines) on individual human chromosomes.

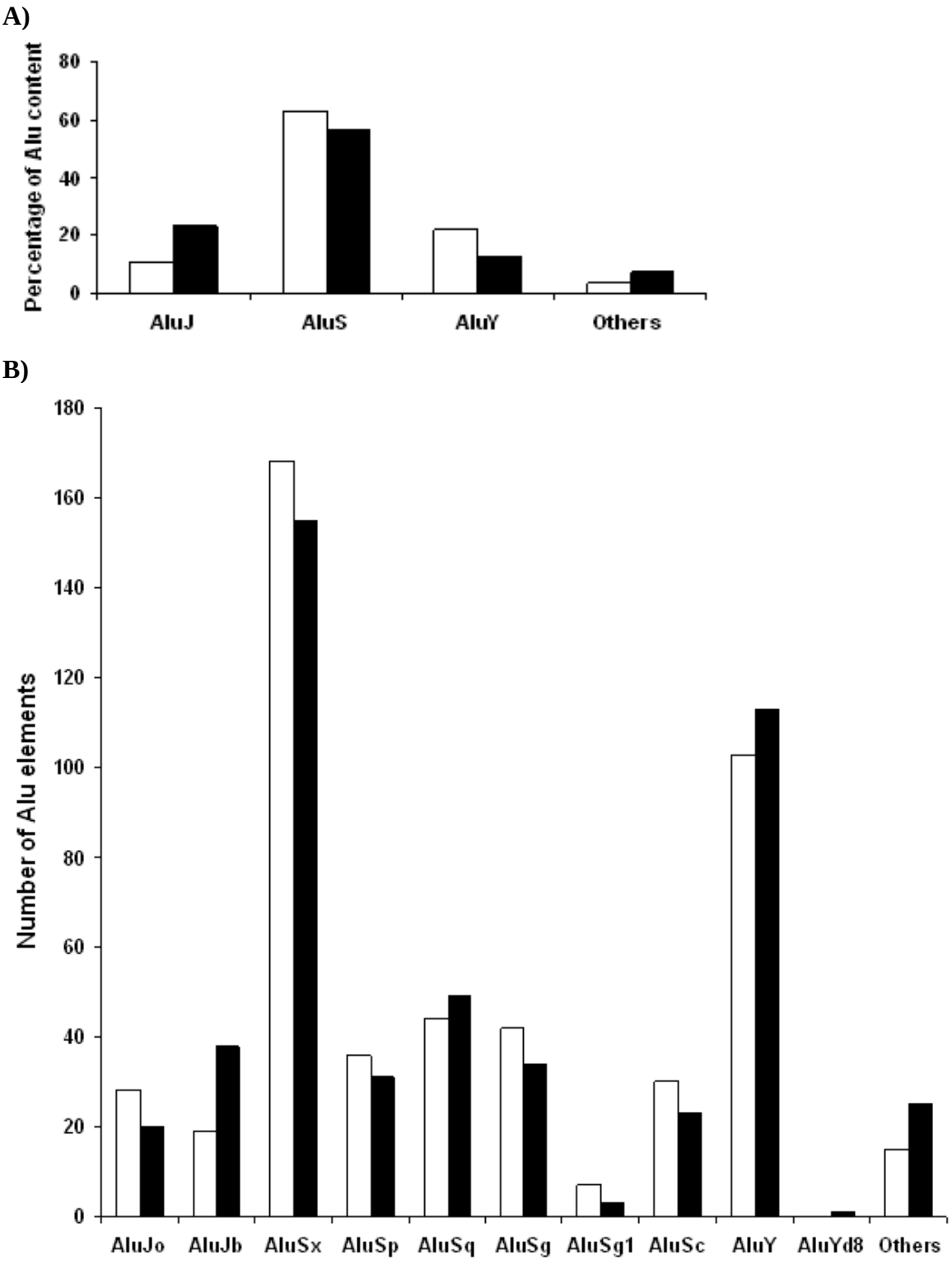


Figure 3.3. *Alu* subfamily composition in ARMD events. (A) Proportion of *Alu* elements involved in ARMD events (unblackened bar) versus total number of *Alu* elements (blackened bar) for each subfamily. (B) Subfamily ratios of upstream and downstream *Alu* elements involved in ARMD events (unblackened and blackened bars, respectively).

older subfamilies (such as *Alu*J) that would have had more time for recombination because of their age. This implies that *Alu* elements throughout the genome have similar chances of recombining with each other, as opposed to a mechanism of preferential recombination between members of an individual subfamily, and that proximity between the elements is the major factor involved in the process. Additional evidence supporting this position comes from the fact that ~40% (197 of 492) of ARMD events result from inter-*Alu* subfamily recombinations. However, within this context, the amount of sequence identity between the two elements at a locus also appears to be proportional to their chances of successful recombination, since young *Alu*Y elements are over-represented at ARMD loci compared with their total number in the genome, whereas the opposite is true for older, highly diverged *Alu*J elements.

The total amount of genomic sequence deleted by this process in the human lineage (i.e., after the human-chimpanzee divergence ~6 million years ago) is estimated to be 396,420 bp. This is probably a conservative estimate, since our comparative analysis of the human and chimpanzee genomes detects ARMD events only between *Alu* elements that were inserted before the human-chimpanzee divergence. Therefore, it would miss ARMD loci involving newly inserted human-specific *Alu* elements (Carter *et al.* 2004; Otieno *et al.* 2004). However, the contribution of human-specific *Alu* elements to ARMD is probably relatively limited, given that there are only ~7000 such insertions (CSAC 2005), as compared with >1 million *Alu* elements shared between the human and chimpanzee genomes.

The ARMDs range in size between 101 and 7255 bp, with an average size of ~806 bp. A histogram of the size frequency distribution of ARMDs reveals a skew towards shorter ARMD sizes, with ~75% (368 of 492) of the deletions shorter than 1 kb (Figure 3.4). Thus, the median ARMD length of 468 bp better represents the most common size category. However, in terms of

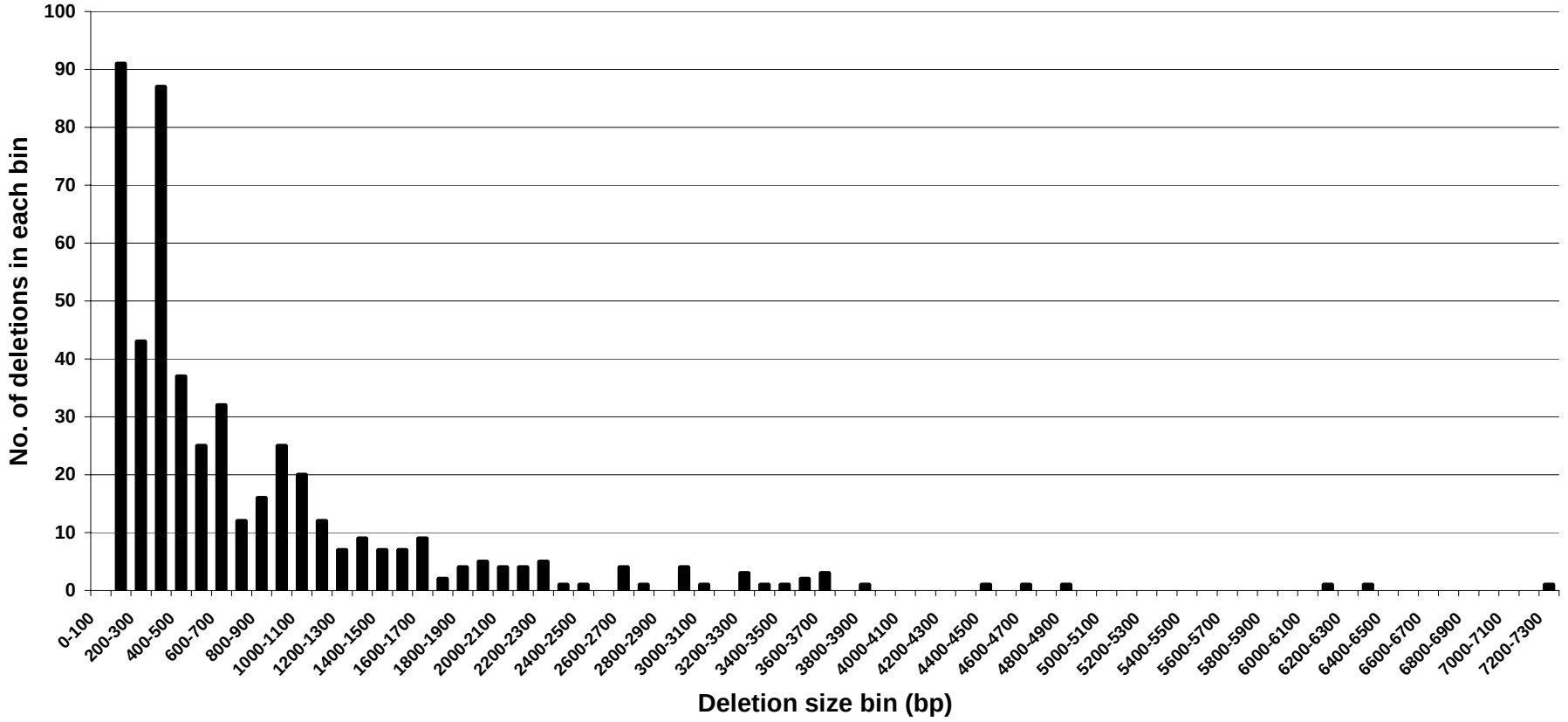


Figure 3.4. Size distribution of human-specific ARMD events, displayed in 100-bp bin sizes.

total genomic sequence deleted, the ~25% ARMD events >1 kb were responsible for ~62% (245,263 of 396,420 bp) of the total sequence deleted. Our computational analyses did not return any ARMD loci with deletions <100 bp. Strictly speaking, *Alu*-*Alu* recombination elements should not cause deletions of <300 bp (i.e., the length of a complete *Alu* element), because, even if the recombining elements were immediately adjacent to each other, this would be the smallest possible amount of sequence deleted. However, the individual left and right monomers of the dimeric *Alu* element can freely exist in the genome, and these types of elements are accounted for in our study. This resulted in the ability of our study to detect deletions smaller than the expected minimum of ~300 bp.

Structural Characteristics of ARMD Events

Pairs of *Alu* elements that recombined to cause human genomic deletions were in parallel orientation in almost all cases (490 of 492). Most probably, this is a direct consequence of the increased length of hybridization available from this arrangement, as the parallel orientation would allow for homology over longer stretches between pairs of *Alu* elements located on the homologous chromosomes during recombination. Analysis of the *Alu* trios at each locus (i.e., two pre-ARMD *Alu* elements in chimpanzee and one postdeletion element in human) suggests four possible recombination mechanisms. Of these, unequal recombination between adjacent *Alu* elements on homologous chromosomes (Figure 3.5A, left panel) accounts for ~74% (366 of 492) of the deletions, whereas the other three putative mechanisms were less frequent (Figure 3.5B-3.5D). Our study captures both intrachromosomal (Figure 3.5A, right panel) and interchromosomal (Figure 3.5A, left panel) recombination-mediated deletions.

For each deletion, we located the points on the *Alu* consensus sequence where the two intact chimpanzee *Alu* elements involved in the recombination were broken and subsequently

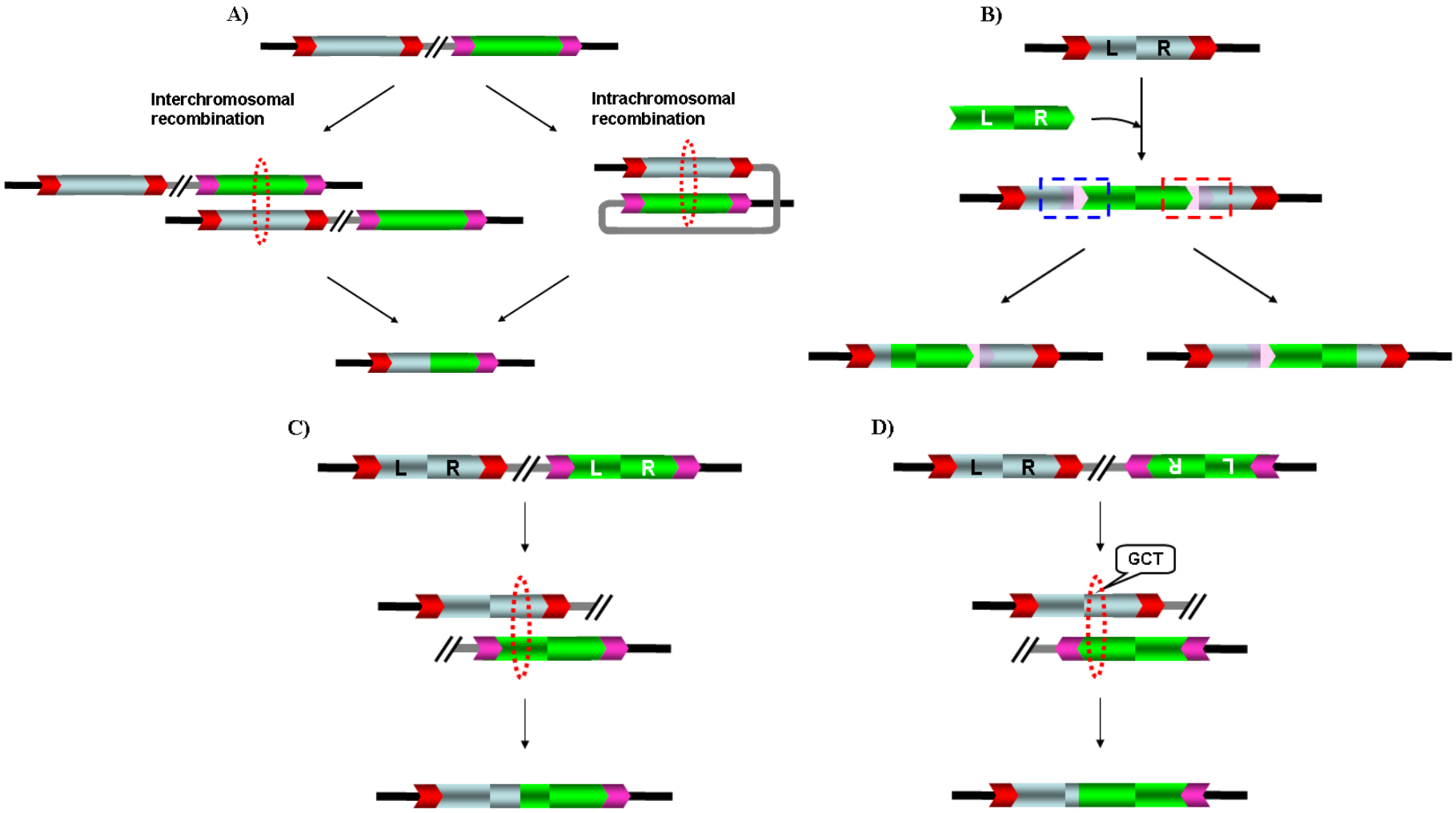


Figure 3.5. Four different types of the recombination between *Alu* elements. Black and gray lines represent flaking and intervening regions, respectively. Dotted red circles denote recombining regions, and red and pink arrows represent TSDs of the two elements, respecitively. (A) Interchromosomal (left) and intrachromosomal (right) recombination between two *Alu* elements (light blue and green). (B) Recombination between two *Alu* elements, one of which previously inserted into the other (L and R indicate left and right *Alu* monomers). (C) Recombination between left and right *Alu* monomers on two different elements. (D) Recombination between oppositely oriented *Alu* elements (only two cases observed).

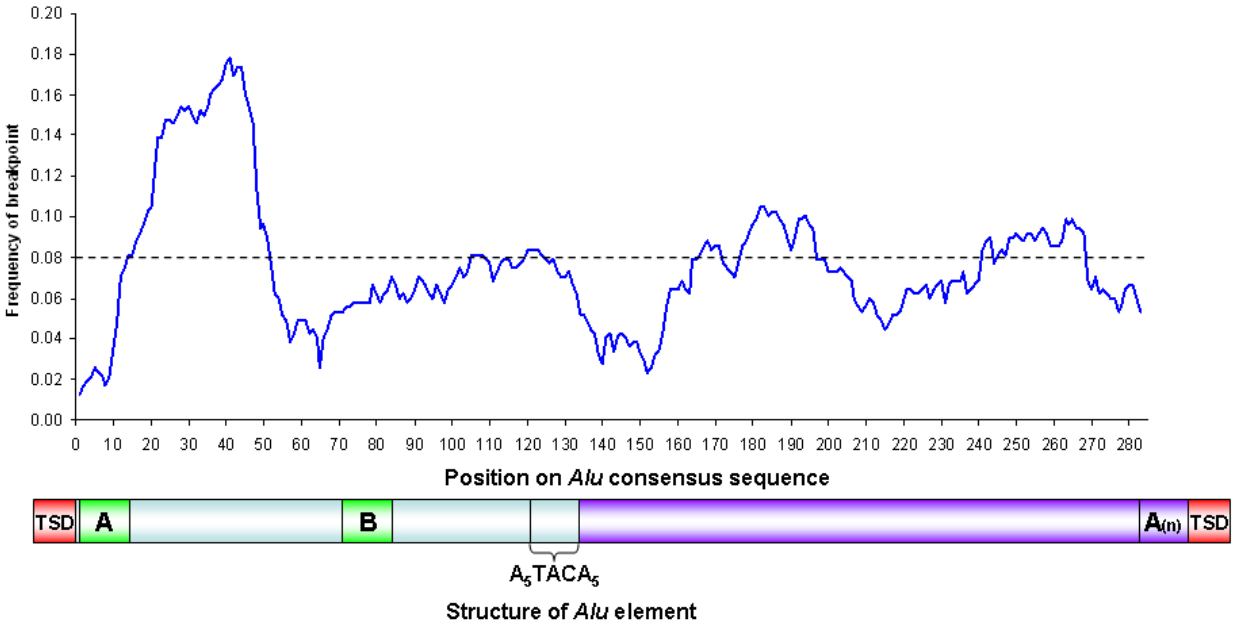


Figure 3.6. Recombination window between *Alu* elements and percentage frequencies of breakage (during recombination) at different positions along an *Alu* consensus sequence. The structure of a typical *Alu* element is shown in the lower panel. The length of the *Alu* consensus sequence is ~282 bp, excluding the 3' poly(A) tail. The element consists of left (light blue) and right (purple) monomers. The left monomer contains an RNA polymerase III promoter (green boxes A and B). TSDs (red boxes), usually between 7-20 bp long, are created at each end during the *Alu* insertion process.

attached to each other to form the resulting single human *Alu* element. Plotting the frequency distribution of recombination breakpoints at different positions on the *Alu* consensus sequence revealed a recombination "hotspot" encompassing positions 21-48 (Figure 3.6), which is consistent with an earlier study based on a smaller dataset (Rudiger *et al.* 1995). To uncover the reasons underlying the observed "adhesive" nature of this part of the *Alu* element, we aligned the consensus sequences of 10 *Alu* subfamilies (*Alu*Jo, *Alu*Jb, *Alu*Sx, *Alu*Sp, *Alu*Sq, *Alu*Sg, *Alu*Sg1, *Alu*Sc, *Alu*Y, and *Alu*Yd8) and analyzed the levels of conservation and GC content of regions that tended to recombine at frequencies exceeding the mean (0.08) across all positions in our ARMD events. This analysis indicated that both parameters were substantially higher in these

regions than in the rest of the *Alu* sequence, with the major inferred recombination hotspot referred to above showing >60% GC (as compared to the ~62.7% average GC content for the 10 *Alu* consensus sequences) and complete conservation across all subfamilies. Although these factors may be responsible for higher recombination frequencies in this region, other reasons are also plausible, such as the location of this stretch near the L1 endonuclease cleavage site at the 5' end of the *Alu* element, which make it closer to putative breakage sites during the recombination process.

Genomic Environment of ARMD Events

Alu elements in the human genome show a preference for high GC content areas, except for the most recently integrated subfamilies (Cordaux et al. 2006a; Lander et al. 2001a). However, since only a fraction (984 of ~1.2 million) of the total number of *Alu* insertions is associated with the ARMD process, it may well be that, in this respect, the deletions themselves behave differently from the *Alu* family as a whole. To characterize the sequence context in which ARMD events occur, we calculated the percentage GC content in 20-kb windows of flanking sequence centered on the ARMD loci. Compared with previous analyses of *Alu* and L1 insertion-mediated (as opposed to postinsertional recombination-mediated) genomic deletions (Callinan et al. 2005; Han et al. 2005), which are preferentially localized in low-GC content neighborhoods (~38% GC), ARMD events tend to occur in high-GC content regions (~45% GC content on average). This is also substantially higher than the ~41% global average GC content of the human genome (Lander et al. 2001a). Since high-GC content areas of the genome also show higher gene density (IHGSC 2004; Lander et al. 2001a), we analyzed 4 Mb (2 Mb in either direction) of sequence flanking ARMD events, for the presence of known and predicted human RefSeq genes. We found the gene density around ARMD events to be, on average, one gene per

66 kb, which, as expected, is higher than the global average gene density (approximately one gene per 150 kb) (IHGSC 2004) and the average gene density in the vicinity of L1 insertion-mediated deletions (approximately one gene per 200 kb) (Han et al. 2005). Thus, ARMD events seem to be concentrated in gene-rich regions of the human genome. The tendency for clustering of ARMD events and genes becomes even more apparent when their densities are plotted side by side on each chromosome (Figure 3.7). Interestingly, the neighboring GC content showed a

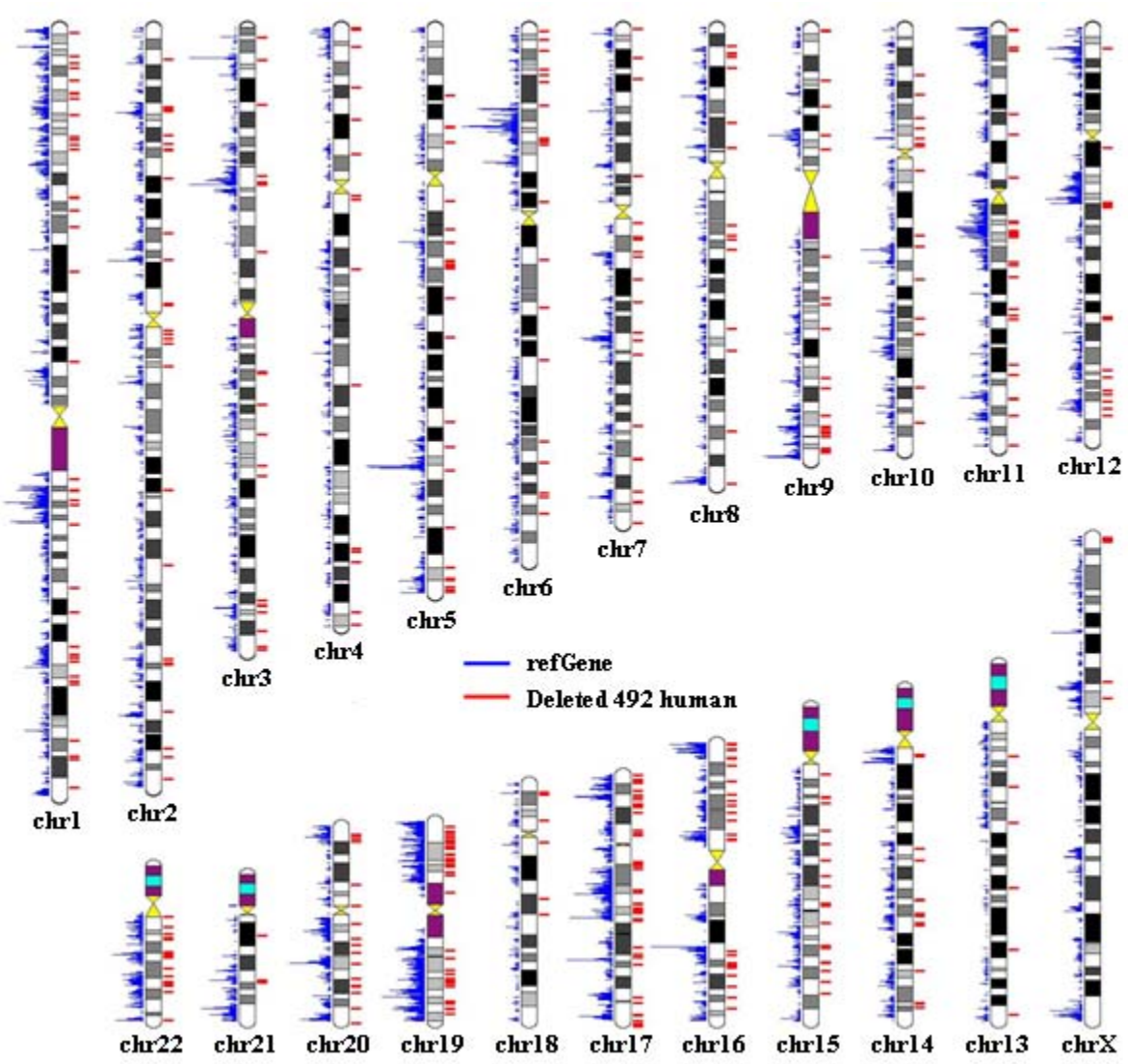


Figure 3.7. Density of ARMD events (red lines) and RefSeq genes (blue lines) on individual human chromosomes.

significant negative correlation with the deletion size (r = -0.17; P < .0001).

About 45% (219 of 492) of ARMD events were located within known or predicted human RefSeq genes, and an additional ~15% (76 of 492) were in intergenic regions of the human genome but were located within predicted chimpanzee genes. Since ≤25% of the human genome represents currently known genes (including both exon and intron sequences) (IHGSC 2004; Sakharkar et al. 2004; Venter et al. 2001), the relative density of ARMD events within genic regions is remarkably high. This would indicate that, *a priori,* the probability of this process interfering with gene function is higher than the two retrotransposon insertion-mediated deletion mechanisms mentioned above. To test this hypothesis, we extracted the ancestral prerecombination sequence at each ARMD locus (i.e., the sequence present in the chimpanzee genome but deleted in the human genome) and analyzed its location in the chimpanzee genome to see whether it mapped to a protein-coding region. In three instances, the ARMD event deleted an entire exon from a gene that is functional in the chimpanzee genome. To confirm that these three ARMD loci did not represent assembly errors, we resequenced them in the human and chimpanzee genomes. One of the three genes, *LOC471177* is a model chimpanzee gene similar to the human *CHRNA9* gene (MIM 605116), a member of the ligand-gated ionic channel family that is associated with cochlea hair cell development (Lustig and Peng 2002). Of the other two, *LOC452742* is similar to the human model gene *LOC440141* (which encodes the mitochondrial ribosomal protein S31), and *LOC471116* encodes a hypothetical protein with a conserved high-molecular weight glutenin subunit.

Characteristics of the Genomic Sequences Lost during ARMD

Previous analyses have suggested that recombination may be responsible for the bias towards high-GC content areas observed for *Alu* elements in the human genome (Brookfield

2001; Hackenberg et al. 2005; Jurka et al. 2004; Lander et al. 2001a). If so, one would expect that ARMD events preferentially remove low-GC content sequence, consequently causing a shift in the opposite direction. However, simulation results revealed that the GC content of both RSNA and RSG (41.9% and 41.4%, respectively) were significantly lower than the ~45.4% GC content of the observed deleted sequences (*P* value < .00001 in both cases). Moreover, the RSNA and RSG *Alu* contents (20.6% and 11.4%, respectively) also had significantly lower values when compared to the *Alu* content of the observed deleted sequences (27.0%; *P* < .0001, compared with both RSNA and RSG). In addition to *Alu* elements, repetitive DNA from elements of other families, for a total of 86,442 bp, was removed by ARMD (Table 3.2).

Table 3.2. Genomic DNA sequences deleted by ARMD

Classification	Amount (bp)
Alu[a]	192,102
MIR	4780
7SL RNA	306
L1	41,491
L2	7312
L3	163
LTR	23,336
MER1	3575
MER2	2555
Other DNA repeat elements	669
Simple repeat	2255
Nonrepetitive DNA	117,876
Total	396,420

[a] Including truncated *Alu* elements.

Discussion

Role of the ARMD Process in Human Genome Evolution

Retrotransposons such as *Alu* elements are associated with size expansion in primate genomes (Liu et al. 2003; Petrov 2001). This is a consequence of their increasing copy number and also an indirect result of their implication in homology-mediated segmental duplications (Bailey *et al.* 2003). For example, the high retrotransposition activity of the *Alu* family in the human lineage has been responsible for the addition of ~2.1 Mb to the human genome within the past ~6 million years (CSAC 2005; Hedges et al. 2004). In this context, our study provides the first comprehensive assessment of a postretrotransposition process that has had an appreciable impact on the dynamics of human genome-size evolution. Previous *in vivo* evolutionary analyses have characterized human and chimpanzee genomic deletions generated on *Alu* and L1 insertion (Callinan et al. 2005; Han et al. 2005). However, the combined extent of human-specific deletion attributable to these mechanisms is an order of magnitude lower than that resulting from ARMD (~30 Kb for *Alu* and L1 insertion-mediated deletions combined, vs. ~400 kb for ARMD alone). The relative amounts of sequence inserted (by *Alu* retrotransposition) and deleted (by ARMD) imply an *Alu*-mediated sequence turnover rate of ~20% (i.e., ~400-kb deleted sequence vs. ~2.1-Mb inserted sequence) in the human genome within the past ~6 million years. This indicates that ARMD is capable of mitigating, at least partially, the increase in genome size caused by new retrotransposon insertions.

The scope of retrotransposon-mediated reduction of genome size further broadens when we consider that L1 elements (another mobile DNA family) are capable of creating deletions by a recombination process analogous to ARMD (Bailey et al. 2003; Burwinkel and Kilimann 1998). The higher average distance between L1 insertions in the human genome (one element

per 6.3 kb) (Lander et al. 2001a) as well as the lower GC content of L1 elements (~43%, excluding the poly(A) tail) (Dombroski *et al.* 1993) may be contributing factors to the paucity of L1-mediated recombination events as compared to ARMD events. Even so, the greater length of L1 elements (~6 kb vs. ~300 bp for *Alu* elements) (Dombroski *et al.* 1993) and their high copy number (~520,000 elements) (Lander et al. 2001a) still indicate that this family may represent another source of retrotransposon recombination-mediated deletions in the human genome. However, a broader comparative genomic study of such retrotransposon recombination-mediated deletion mechanisms in both the human and chimpanzee lineages is needed before the comprehensive role of transposable elements in primate genome-size evolution can be determined. In this respect, at least in the case of plants, studies have already shown that the genome of *Arabidopsis thaliana* uses recombination-mediated deletion to counterbalance genome expansion, which may be one of the reasons for its remarkably compact size (Devos *et al.* 2002).

Recent analyses of human-genome variation have emphasized the importance of deletions in creating genetic diversity among humans (Conrad et al. 2006; Hinds et al. 2006; Iafrate et al. 2004; McCarroll et al. 2006). Our results offer insight into one of the mechanisms that may contribute to the creation of such deletions. Interestingly, the majority of the deletion variants identified in the recent studies cited above (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2005) are polymorphic between human individuals or populations. Although their contribution to between-individual genetic diversity is undisputed, the persistence of these deletions over evolutionary time cannot be taken for granted. By contrast, the deletions reported in our study have a low polymorphism rate (15%) among the 80 diverse human genomes we genotyped. This may represent the difference in the comparative timescales of these between-

human genomic deletion variants (Conrad *et al.* 2006; Hinds *et al.* 2006) and our human-chimpanzee comparison. In an earlier analysis (Han et al. 2005), we showed that only a fraction of the deletions caused by *in vitro* L1 retrotransposition (Gilbert et al. 2002; Gilbert et al. 2005; Symer et al. 2002) persist in the human genome over evolutionary time. Additionally, comparative genomic studies across a range of organisms indicate that genomic deletions that ultimately reach fixation tend to be smaller than those detected before any selective force operates (i.e., in cell culture analyses) (Gregory 2004). Analogous to this situation, ARMD events (which had a median length of 468 bp) were, in general, smaller than the deletion variants characterized by the recent studies of human-genome variation, which had a range of 1-745 kb (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2005). Since our study focuses on a longer evolutionary time scale and would preferentially capture those ARMD events that have not been selected against, it is possible that the deletions we detected represent the smaller evolutionary remainder of a group of older and perhaps larger deletions.

ARMD as an Agent in Human-Chimpanzee Divergence

The human and chimpanzee genomes are characterized by only ~1.4% divergence at the nucleotide-sequence level (CSAC 2005; Ebersberger et al. 2002; Newman et al. 2005; Watanabe et al. 2004). With the completion of the draft chimpanzee genome, the focus has shifted to identifying differences rather than locating similarities. Regarding actual genetic change, although a comprehensive assessment of protein-coding portions of the chimpanzee genome is not yet available, functional classes of genes that are under accelerated evolution in one lineage or the other have been characterized by recent studies (Clark *et al.* 2003; Dorus *et al.* 2004).

In the context of possible events that have altered gene structure or expression between the human and chimpanzee lineages, our study illustrates almost 300 lineage-specific deletions

within protein-coding human or chimpanzee RefSeq genes; it is conceivable that at least some of these ARMD events contributed to phenotypic divergence. Gene shuffling by recombination between *Alu* elements has already been reported in the human genome (Babcock *et al.* 2003). Furthermore, in at least two documented instances, *Alu* elements have caused hominoid lineage-specific exon deletions in functional genes: through an insertion-mediated deletion in the human *CMAH* gene (Hayakawa *et al.* 2001) and through ARMD in the human *ELN* gene (Szabo *et al.* 1999). In the present study, we show three additional instances in which ARMD has caused the loss of an exon in a human gene, as compared to its chimpanzee ortholog. Of particular interest is the deletion of the fourth exon in the predicted chimpanzee gene *LOC471177,* which is orthologous to the human *CHRNA9* gene. In the human lineage, CHRNA9 is an ionotropic receptor with a probable role in the modulation of auditory stimuli (Glowatzki and Fuchs 2000; Lustig and Peng 2002). Modifications in the function of this gene may lead to a reduction in basiliar membrane movement and thus affect the dynamic range of hearing. Although the characterization of the actual gene expression pathways that underlie the differences of humans and chimpanzees has just begun, preliminary data suggest that differences in auditory genes may comprise a subset of the total change (Clark *et al.* 2003). This is reflected in the fact that the tonal range of normal human speech is probably outside the optimal reception of the chimpanzee auditory system (Martinez *et al.* 2004). Thus, it is conceivable that *CHRNA9* is a member of the group of genes (such as *FOXP2* and *TECTA*) that may be responsible for the unique auditory and olfactory traits that distinguish humans and chimpanzees (Clark et al. 2003; Enard et al. 2002). Even excluding the three ARMD events listed above that deleted exons, 292 other events located within genes have deleted 229,205 bp of intronic sequence. Although further analysis will be

required for conclusive assignment of specific roles, if any, to the deleted intronic sequencse, it is possible that some of them may be associated with alteration of splicing patterns.

Does ARMD Play a Role in Modifying *Alu* Distribution?

Recently integrated or young *Alu* elements are inserted relatively randomly in the genome; by contrast, older *Alu* elements are preferentially found in GC-rich areas of the genome (Cordaux et al. 2006a; Lander et al. 2001a). Both selective and neutral explanations have been offered for this uneven genomic distribution of *Alu* elements. However, a selective process (Lander et al. 2001a) is inconsistent with polymorphism patterns of recently integrated *Alu* elements (Cordaux et al. 2006a). An alternative neutral explanation for the enrichment of *Alu* elements in GC-rich regions over time involves their preferential loss from GC-poor regions (Brookfield 2001; Hackenberg et al. 2005; Jurka et al. 2004; Lander et al. 2001a), a process that might be influenced by ARMD.

However, the high GC content of deleted sequences, along with the preferential occurrence of ARMD events in GC-rich regions, argues against this possibility. To result in the *Alu* distribution shift, the deletions would need to be much larger in GC-poor than in GC-rich regions (Cordaux et al. 2006a). Consistent with this hypothesis, our results indicate that ARMD size is negatively correlated with GC content. However, although ARMD events are significantly larger in GC-poor (i.e., <41% genome average) than in GC-rich (i.e., >41% genome average) regions (~1100 vs. ~700 bp; *t* test, $P = .0007$), three times as many ARMD events occurred in GC-rich as in GC-poor regions (369 vs. 123). Consequently, the net amount of sequence deleted from GC-poor regions is half that of GC-rich regions (~135 kb vs. ~261 kb). Given that GC-poor regions encompass ~58% of the genome (Lander et al. 2001a), it is unlikely that ARMD has played a substantial role in mediating the shift in the *Alu* distribution towards heavy isochors

(CSAC 2005). Nevertheless, other types of deletions could contribute more significantly to the yet-unexplained *Alu* genomic distribution shift.

Interestingly, the results from the simulations we performed suggest that sequences deleted through ARMD contain a statistically significant excess of *Alu* elements. This implies that the ARMD process may contribute to effective removal of *Alu* elements from regions in which they have reached high densities. Given the fact that abnormally high *Alu* density within a particular genomic region would also make it prone to recombination-mediated deletions, this result may reflect a selective force that counteracts the deletion process.

A Potential Mechanism of Double-Strand Break (DSB) Repair

Previous analyses have demonstrated the ability of both LTR and non-LTR retrotransposons to cause DSBs in genomic DNA (Gasior et al. 2006; Zimmerly et al. 1995). In particular, the role of the L1 family in the creation and subsequent resolution of DSBs has been extensively analyzed (Gilbert *et al.* 2005). *In vitro*, cell-culture studies have shown that homology-directed repair is a major mechanism for patching such breaks and that recombination between repetitive elements is one possible pathway for this process (Richardson and Jasin 2000). Recombination rates are highly increased on artificially induced DSBs in cultured cells, which further implicates this mechanism in "tying up the loose ends" at potentially deleterious DSB loci (Liang *et al.* 1998).

In vitro, a 3:1 excess of recombination deletions versus conservative noncrossover situations was detected in a study of homology-mediated repair at a single predefined DSB locus (Liang *et al.* 1998). In this context, some of the loci in our study may represent instances of homology-mediated DSB repair, in which the presence of highly conserved *Alu* sequences on both sides of the break has facilitated its patching. This would be particularly true for loci at

which the deletion would otherwise be selectively neutral, since the act of having repaired a potentially lethal DSB would give it an instant advantage, if only for propagation to the immediately next generation.

Conclusion

As high-throughput sequencing techniques become more advanced, the focus of evolutionary studies is shifting more towards genomewide analyses. Our study represents such a situation: we have comprehensively analyzed a major deletion mechanism in the human genome that was previously known only as a result of mutations in isolated disease-causing loci. In view of the fact that deletions are being recognized as an important class of genetic variants that contribute to human diversity and evolution (Conrad *et al.* 2006; Hinds *et al.* 2006; McCarroll *et al.* 2006), ARMD represents one of the major mechanisms for generating such deletions in humans. Moreover, the frequent occurrence of ARMD in gene-rich regions of the genome demonstrates the importance of this process in both biomedical and evolutionary studies. Overall, our results open the field to further studies of deletions caused by recombination between mobile elements and demonstrate one of the possible ways by which the human lineage may have developed a set of unique genetic traits.

Materials and Methods

Computational Data Mining for Identifying Candidate ARMD Loci

We extracted 400 bp of 5' and 3' genomic sequence flanking all human *Alu* elements (Figure 3.8). Next, we joined the two 400-bp stretches to form a single sequence (the "query"). For each query, the best match in the reference chimpanzee genome (PanTro1 [November 2003 freeze]) was identified. Then, the sequence stretch in the chimpanzee genome between the two regions that aligned with the two-400 bp halves of the query (the "hit") was extracted and

aligned with the human *Alu* sequence initially used to design the query (the "query *Alu*"), by use of a local installation of the National Center for Biotechnology Information Blast 2 Sequences Bl2seq utility. Following are the possible alignment results for each sequence pair (see corresponding diagrams in Figure 3.8).

A. There is no match. In this case, an *Alu* insertion-mediated deletion has occurred in the human genome at that locus.

B. There is only one alignment block, and:

B.1. The hit is identical to the query *Alu*. This is shared ancestry of an *Alu* insertion.

B.2. The hit is longer than the query *Alu*, and the extra sequence is entirely composed of a poly(A) tract downstream of the *Alu* sequence. This is a case of extension of the *Alu* poly(A) tail.

B.3. The hit consists of the query *Alu* plus some extra non-poly(A) sequence, and:

B.3a. The extra, non-poly(A) sequence is downstream of the poly(A) tail. This could be a gene conversion event in the chimpanzee genome.

B.3b. The extra, non-poly(A) sequence is upstream of the query *Alu* element or there is extra sequence at both ends. This is a possible *Alu* insertion-mediated deletion event in the human genome.

C. There is more than one alignment blocks, and:

C.1. The beginning and end of the hit match the query *Alu* and the hit is at least 100 bp longer than the query *Alu* sequence (since this size would approximate the expected lower ARMD size limit). This is a candidate ARMD event in the human genome.

C.2. At least one end of the hit has no match to the query *Alu*. This is another possible case for an *Alu* insertion-mediated deletion in the human genome.

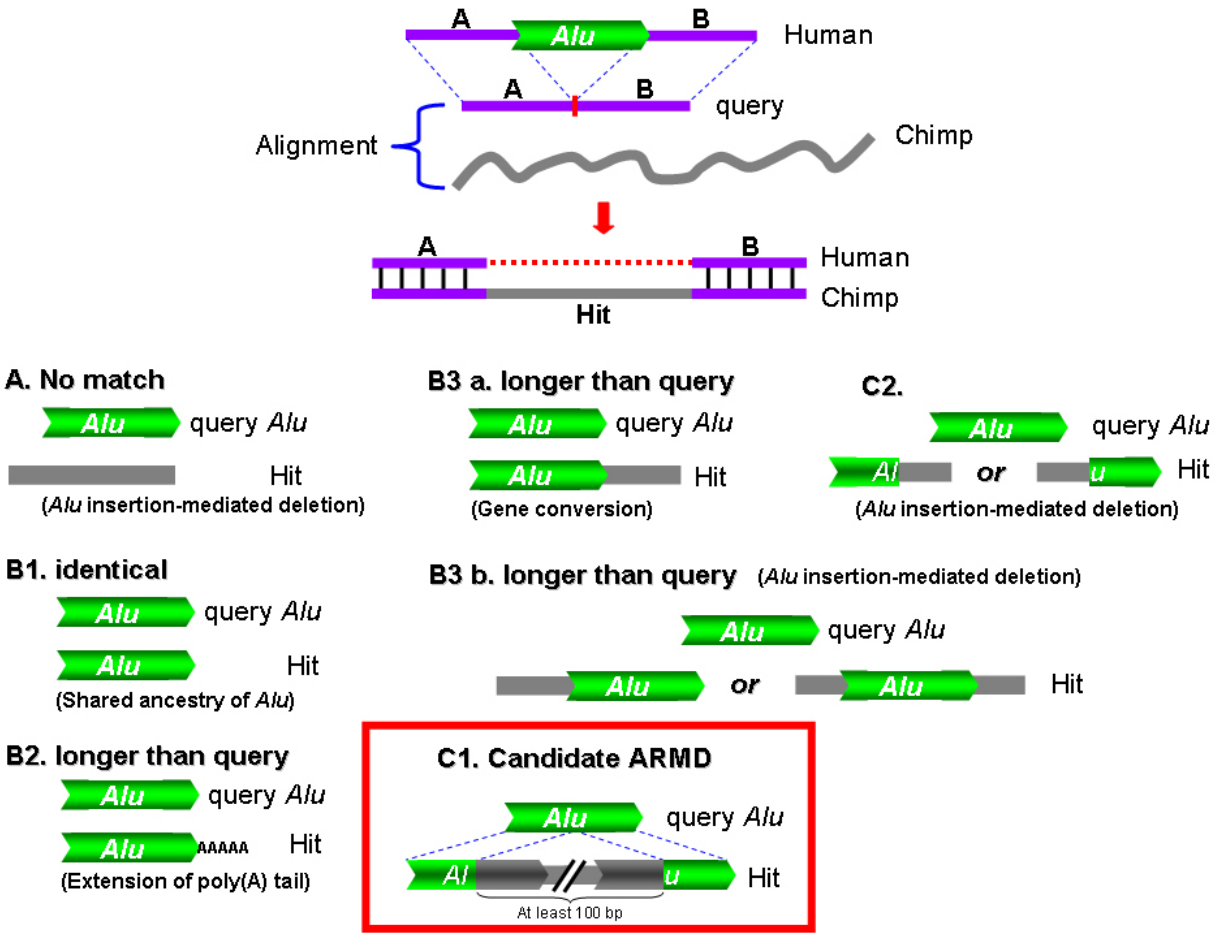


Figure 3.8. Computational data mining for human lineage-specific ARMD loci.
(A) No match between query *Alu* and hit (possible *Alu* insertion-mediated deletion). (B.1) Query *Alu* and hit are identical (shared ancestry of an *Alu* insertion). (B.2) Hit is longer than query *Alu* and the extra sequence is a poly(A) tract downstream of the query *Alu* (extension of the *Alu* poly(A) tail). (B.3) Hit consists of query *Alu* plus extra non-poly(A) sequence, and the following. (B.3a) Extra, non-poly(A) sequence is downstream of the query *Alu* poly(A) tail (may be gene conversion event in the chimpanzee genome). (B.3b) Extra, non-poly(A) sequence is upstream of the query *Alu* element or there is extra sequence at both ends (possible *Alu* insertion-mediated deletion event). (C.1) Beginning and end of the hit match query *Alu* and the hit is at least 100 bp longer than query *Alu* (candidate human lineage-specific ARMD event). (C.2) At least one end of the hit has no match to query *Alu* (possible *Alu* insertion-mediated deletion).

We retained all loci matching case C.1 as pairs of FASTA files (i.e, the orthologous human and chimpanzee sequences). Each human sequence contained the query *Alu* and its 400-bp flanking sequences on each side, and each chimpanzee sequence contained the entire hit that aligned with the query flanking sequences. All candidate ARMD loci were then manually

inspected and, if necessary, verified by wet bench (PCR) analysis. Orthologous human and chimpanzee sequences for each locus are available from the "Publications" section of the Batzer Laboratory Web site.

Inspection of Target Site Duplications (TSDs)

A typical *Alu* insertion is flanked on both sides by identical (or nearly perfect) short, direct repeats (7-20 bp) termed "target-site duplications" (TSDs) (Deininger and Batzer 2002). The single *Alu* element remaining at a human candidate ARMD locus is characterized by the apparent absence of TSDs, since it is composed of fragments from a pair of *Alu* elements with mutually different TSDs, situated at the orthologous ancestral locus (which persists in the chimpanzee genome). This hallmark of the ARMD process offers a direct means of confirming the "chimeric" origin of the human *Alu* element at a deletion locus. Using this property as our basis for verification, we manually inspected all candidate loci returned by the computational analysis. In an unambiguous ARMD event, the TSDs of the two *Alu* elements immediately upstream and downstream of the deleted portion in the chimpanzee genome were perfect matches with the 5' and 3' TSDs, respectively, of the orthologous single human *Alu* element. In the next possible scenario, the sequence on any one side of the human *Alu* (upstream or downstream) matched the TSDs of the chimpanzee element on the corresponding side, but the other chimpanzee *Alu* element itself lacked TSDs. However, the sequence immediately flanking this element on the side opposite to the deletion was identical in both human and chimpanzee. In both these cases, we accepted the computational detection as a valid ARMD locus. At loci that showed slight deviations in the sequence architecture from the unambiguous ARMD structures described above (which raise the possibility that one of the two chimpanzee *Alu* elements might be a chimpanzee-specific *Alu* insertion, as opposed to a human-specific ARMD event), we

designed oligonucleotide primers in the nonrepetitive sequences flanking the *Alu* elements in the chimpanzee genome and we experimentally confirmed by PCR (and, where required, by DNA sequencing) that the deletion did exist and was specific to the human genome.

As an additional step to verify the potential ARMD loci that we accepted/rejected based solely on computational identification, we randomly chose two sets of 25 such insertions and deletions and verified them by PCR. Accuracy rates for putative deletion and insertion loci were 100% and 96%, respectively (4% of putative insertions comprising the error were all deletions), confirming the validity of our approach.

PCR Amplification and DNA Sequence Analysis of ARMD Loci

We designed oligonucleotide primers using Primer3 software. Detailed information for each locus including primer sequences, annealing temperature and PCR product sizes is available from the "Publications" section of the Batzer Laboratory Web site.

PCR amplification of each locus was performed in 25 μl reactions with 10-50 ng genomic DNA, 200 nM of each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM MgCl2, 10 mM Tris-HCl (pH 8.4), and 2.5 units *Taq* DNA polymerase. The conditions for the PCR were an initial denaturation step of 94°C for 4 min; followed by 32 cycles of 1 min of denaturation at 94°C, 1 min of annealing at optimal annealing temperature, and 1 min of extension at 72°C; followed by a final extension step at 72°C for 10 min. PCR amplicons were separated on 2% agarose gels, were stained with ethidium bromide, and were visualized using UV fluorescence.

Individual PCR products were purified from the gels with Wizard gel purification kits (Promega) and were cloned into vectors by use of TOPO-TA Cloning kits (Invitrogen). For each sample, three colonies were randomly selected and sequenced on an Applied Biosystems

ABI3130*XL* automated DNA sequencer. Each clone was sequenced in both directions with use of M13 forward and reverse primers. The sequence tracks were analyzed using the Seqman program in the DNASTAR suite and were aligned using BioEdit sequence alignment software. Gorilla and orangutan sequences generated during the course of this study have been submitted to GenBank under accession numbers DQ363502-363524.

Loci verified by PCR were screened on a panel of five primate species, including *Homo sapiens* (HeLa; cell line ATCC CCL-2), *P. troglodytes* (common chimpanzee; cell line AG06939B), *Pan paniscus* (bonobo or pygmy chimpanzee; cell line AG05253B), *Gorilla gorilla* (Western lowland gorilla; cell line AG05251) and *Pongo pygmaeus* (orangutan; cell line ATCC CR6301). To evaluate polymorphism rates, we amplified 50 randomly picked ARMD loci on a panel of genomic DNA, from 80 human individuals (20 from each of four populations: African American, South American, European, and Asian) that was available from previous studies in our lab.

Monte Carlo Simulations of GC and *Alu* Content

To test whether the GC and *Alu* contents of the sequences deleted through ARMD differed statistically from the rest of the genome, we performed Monte Carlo simulations comparing the observed deletions to two other sets of sequences. Both these sets comprised randomly extracted sequences equal in number to the observed deletions (492) and mimicked the observed size distribution of ARMD events. The first set was extracted from the regions immediately adjacent to randomly picked *Alu* elements annotated in the reference human genome sequence (called "RSNA"). The second set comprised sequences randomly extracted from the entire genome sequence, with no additional parameters incorporated (called "RSG"). We used 5000 randomized replicates of both sets. For both observed and simulated sets of

sequences, we calculated GC content using in-house Perl scripts, whereas the *Alu* content was analyzed using a locally installed copy of the RepeatMasker Web server. Additionally, to make our estimate of observed percentage *Alu* content conservative, we trimmed the deleted sequence at each locus to remove remaining fragments of the two *Alu* elements that caused the ARMD event.

Statistical significances of the differences in GC and *Alu* content were based on Z scores obtained by comparing observed values (from the actual set of deleted sequences) with the mean value obtained from the 5000 randomly extracted sequence sets (Hamaker 1978). All computer programs used are available from the authors on request.

References

Babcock, M., A. Pavlicek, E. Spiteri, C.D. Kashork, I. Ioshikhes, L.G. Shaffer, J. Jurka, and B.E. Morrow. 2003. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res* **13:** 2519-2532.

Bailey, J.A., G. Liu, and E.E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Brookfield, J.F. 2001. Selection on Alu sequences? *Curr Biol* **11:** R900-901.

Burwinkel, B. and M.W. Kilimann. 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277:** 513-517.

Callinan, P.A., J. Wang, S.W. Herke, R.K. Garber, P. Liang, and M.A. Batzer. 2005. Alu Retrotransposition-mediated Deletion. *J Mol Biol* **348:** 791-800.

Carter, A.B., A.H. Salem, D.J. Hedges, C.N. Keegan, B. Kimball, J.A. Walker, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2004. Genome-wide analysis of the human Alu Yb-lineage. *Hum Genomics* **1:** 167-178.

Clark, A.G., S. Glanowski, R. Nielsen, P.D. Thomas, A. Kejariwal, M.A. Todd, D.M. Tanenbaum, D. Civello, F. Lu, B. Murphy et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302:** 1960-1963.

Conrad, D.F., T.D. Andrews, N.P. Carter, M.E. Hurles, and J.K. Pritchard. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38:** 75-81.

Cordaux, R., J. Lee, L. Dinoso, and M.A. Batzer. 2006. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene* **373:** 138-144.

CSAC. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67:** 183-193.

Deininger, P.L. and M.A. Batzer. 2002. Mammalian retroelements. *Genome Res* **12:** 1455-1465.

Devos, K.M., J.K. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12:** 1075-1079.

Dombroski, B.A., A.F. Scott, and H.H. Kazazian, Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc Natl Acad Sci U S A* **90:** 6513-6517.

Dorus, S., E.J. Vallender, P.D. Evans, J.R. Anderson, S.L. Gilbert, M. Mahowald, G.J. Wyckoff, C.M. Malcom, and B.T. Lahn. 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell* **119:** 1027-1040.

Ebersberger, I., D. Metzler, C. Schwarz, and S. Paabo. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* **70:** 1490-1497.

Enard, W., M. Przeworski, S.E. Fisher, C.S. Lai, V. Wiebe, T. Kitano, A.P. Monaco, and S. Paabo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418:** 869-872.

Gasior, S.L., T.P. Wakeman, B. Xu, and P.L. Deininger. 2006. The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *J Mol Biol*.

Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Gilbert, N., S. Lutz, T.A. Morrish, and J.V. Moran. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25:** 7780-7795.

Glowatzki, E. and P.A. Fuchs. 2000. Cholinergic synaptic inhibition of inner hair cells in the neonatal mammalian cochlea. *Science* **288:** 2366-2368.

Gregory, T.R. 2004. Insertion-deletion biases and the evolution of genome size. *Gene* **324:** 15-34.

Hackenberg, M., P. Bernaola-Galvan, P. Carpena, and J.L. Oliver. 2005. The biased distribution of alus in human isochores might be driven by recombination. *J Mol Evol* **60:** 365-377.

Hamaker, H.C. 1978. Approximating the cumulative normal distribution and its inverse. *Appl. Statist.* **27:** 76-77.

Han, K., S.K. Sen, J. Wang, P.A. Callinan, J. Lee, R. Cordaux, P. Liang, and M.A. Batzer. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* **33:** 4040-4052.

Hattori, Y., N. Okayama, Y. Ohba, Y. Yamashiro, K. Yamamoto, I. Tsukimoto, and M. Kohakura. 1999. The precise breakpoints of a Filipino-type alpha-thalassemia-1 deletion found in two Japanese. *Hemoglobin* **23:** 239-248.

Hayakawa, T., Y. Satta, P. Gagneux, A. Varki, and N. Takahata. 2001. Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci U S A* **98:** 11399-11404.

Hedges, D.J., P.A. Callinan, R. Cordaux, J. Xing, E. Barnes, and M.A. Batzer. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* **14:** 1068-1075.

Hinds, D.A., A.P. Kloek, M. Jen, X. Chen, and K.A. Frazer. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38:** 82-85.

Huang, L.S., M.E. Ripps, S.H. Korman, R.J. Deckelbaum, and J.L. Breslow. 1989. Hypobetalipoproteinemia due to an apolipoprotein B gene exon 21 deletion derived by Alu-Alu recombination. *J Biol Chem* **264:** 11394-11400.

Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949-951.

IHGSC. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Inoue, K. and J.R. Lupski. 2002. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* **3:** 199-242.

Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101:** 1268-1272.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Levran, O., N.A. Doggett, and A.D. Auerbach. 1998. Identification of Alu-mediated deletions in the Fanconi anemia gene FAA. *Hum Mutat* **12:** 145-152.

Liang, F., M. Han, P.J. Romanienko, and M. Jasin. 1998. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc Natl Acad Sci U S A* **95:** 5172-5177.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.

Lustig, L.R. and H. Peng. 2002. Chromosome location and characterization of the human nicotinic acetylcholine receptor subunit alpha (alpha) 9 (CHRNA9) gene. *Cytogenet Genome Res* **98:** 154-159.

Marshall, B., G. Isidro, and M.G. Boavida. 1996. Insertion of a short Alu sequence into the hMSH2 gene following a double cross over next to sequences with chi homology. *Gene* **174:** 175-179.

Martinez, I., M. Rosa, J.L. Arsuaga, P. Jarabo, R. Quam, C. Lorenzo, A. Gracia, J.M. Carretero, J.M. Bermudez de Castro, and E. Carbonell. 2004. Auditory capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *Proc Natl Acad Sci U S A* **101:** 9976-9981.

McCarroll, S.A., T.N. Hadnott, G.H. Perry, P.C. Sabeti, M.C. Zody, J.C. Barrett, S. Dallaire, S.B. Gabriel, C. Lee, M.J. Daly et al. 2005. Common deletion polymorphisms in the human genome. *Nat Genet* **doi:10.1038/ng1696**.

McCarroll, S.A., T.N. Hadnott, G.H. Perry, P.C. Sabeti, M.C. Zody, J.C. Barrett, S. Dallaire, S.B. Gabriel, C. Lee, M.J. Daly et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38:** 86-92.

Miyamoto, M.M., J.L. Slightom, and M. Goodman. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* **238:** 369-373.

Myerowitz, R. and N.D. Hogikyan. 1987. A deletion involving Alu sequences in the beta-hexosaminidase alpha-chain gene of French Canadians with Tay-Sachs disease. *J Biol Chem* **262:** 15396-15399.

Newman, T.L., E. Tuzun, V.A. Morrison, K.E. Hayden, M. Ventura, S.D. McGrath, M. Rocchi, and E.E. Eichler. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15:** 1344-1356.

Otieno, A.C., A.B. Carter, D.J. Hedges, J.A. Walker, D.A. Ray, R.K. Garber, B.A. Anders, N. Stoilova, M.E. Laborde, J.D. Fowlkes et al. 2004. Analysis of the Human Alu Ya-lineage. *J Mol Biol* **342:** 109-118.

Petrov, D.A. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* **17:** 23-28.

Richardson, C. and M. Jasin. 2000. Coupled homologous and nonhomologous repair of a double-strand break preserves genomic integrity in mammalian cells. *Mol Cell Biol* **20:** 9068-9075.

Rohlfs, E.M., N. Puget, M.L. Graham, B.L. Weber, J.E. Garber, C. Skrzynia, J.L. Halperin, G.M. Lenoir, L.M. Silverman, and S. Mazoyer. 2000. An Alu-mediated 7.1 kb deletion of BRCA1 exons 8 and 9 in breast and ovarian cancer families that results in alternative splicing of exon 10. *Genes Chromosomes Cancer* **28:** 300-307.

Rothberg, P.G., S. Ponnuru, D. Baker, J.F. Bradley, A.I. Freeman, G.W. Cibis, D.J. Harris, and D.P. Heruth. 1997. A deletion polymorphism due to Alu-Alu recombination in intron 2 of the retinoblastoma gene: association with human gliomas. *Mol Carcinog* **19:** 69-73.

Rudiger, N.S., N. Gregersen, and M.C. Kielland-Brandt. 1995. One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Res* **23:** 256-260.

Sakharkar, M.K., V.T. Chow, and P. Kangueane. 2004. Distributions of exons and introns in the human genome. *In Silico Biol* **4:** 387-393.

Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Szabo, Z., S.A. Levi-Minzi, A.M. Christiano, C. Struminger, M. Stoneking, M.A. Batzer, and C.D. Boyd. 1999. Sequential loss of two neighboring exons of the tropoelastin gene during primate evolution. *J Mol Evol* **49:** 664-671.

Tvrdik, T., S. Marcus, S.M. Hou, S. Falt, P. Noori, N. Podlutskaja, F. Hanefeld, P. Stromme, and B. Lambert. 1998. Molecular characterization of two deletion events involving Alu-sequences, one novel base substitution and two tentative hotspot mutations in the hypoxanthine phosphoribosyltransferase (HPRT) gene in five patients with Lesch-Nyhan syndrome. *Hum Genet* **103:** 311-318.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Watanabe, H., A. Fujiyama, M. Hattori, T.D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382-388.

Wildman, D.E., M. Uddin, G. Liu, L.I. Grossman, and M. Goodman. 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. *Proc Natl Acad Sci U S A* **100:** 7181-7188.

Zimmerly, S., H. Guo, P.S. Perlman, and A.M. Lambowitz. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82:** 545-554.

CHAPTER FOUR:

ENDONUCLEASE-INDEPENDENT INSERTION PROVIDES AN ALTERNATIVE PATHWAY FOR L1 RETROTRANSPOSITION IN THE HUMAN GENOME*

*Reprinted by permission of Nucleic Acids Research

Introduction

Long Interspersed Element-1 (LINE-1 or L1) is an ubiquitous retrotransposon family in the human genome, with ~520,000 insertions comprising ~17% of total genomic sequence (Lander et al. 2001b; Smit 1996). A full-length L1 element is ~6 Kb long and contains two open reading frames (ORF1 and ORF2) (Moran and Gilbert 2002). While ORF1 encodes an RNA-binding protein with nucleic acid chaperone activity (Martin 2006), ORF2 encodes for endonuclease (EN) and reverse transcriptase (RT) activities (Feng et al. 1996; Mathias et al. 1991), and both ORFs are required for L1 retrotransposition (Moran et al. 1996; Wei et al. 2001). In addition to insertional mutagenesis (Gilbert et al. 2002; Han et al. 2005; Symer et al. 2002), L1 elements have also been associated with exon shuffling, creation of deletions through unequal homologous recombination, and intra-chromosomal and inter-chromosomal translocation of genomic sequence (Burwinkel and Kilimann 1998; Moran et al. 1999; Pickeral et al. 2000). As such, the dynamic nature of L1 elements makes them important agents of genomic rearrangement (Kazazian 2000; Kazazian and Moran 1998).

The currently accepted model for genomic integration of L1 elements is termed target-site primed reverse transcription (TPRT) (Cost et al. 2002; Luan et al. 1993)(Fig. 4.1). During TPRT, the L1 EN cleaves one strand of the target DNA at a motif approaching the consensus 5'-TTTT/A-3' (where / denotes the cleavage site), producing a free 3'-hydroxyl (Cost and Boeke 1998; Feng et al. 1996). Next, the L1 RNA anneals to the nick site using its 3' poly (A) tail, and the L1 RT initiates reverse transcription using the L1 RNA as a template. Cleavage of the second DNA strand by the L1 EN usually occurs 7-20 base pairs downstream of the initial nicking site, creating staggered breaks in the target DNA that are later filled in to form direct repeats flanking the newly inserted element (termed target site duplications or TSDs)

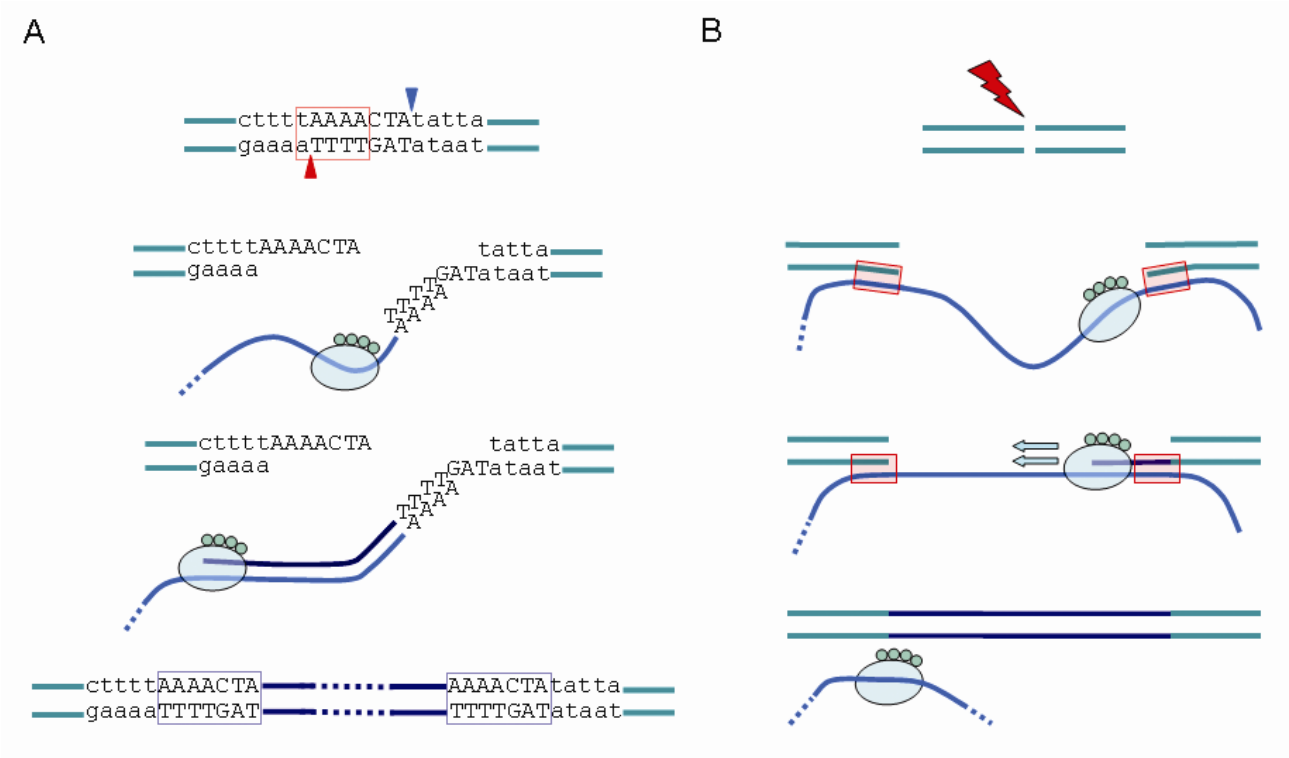


Figure 4.1: Comparison of TPRT and NCLI L1 insertions.
(A) Classical TPRT-mediated L1 insertion in the human genome. First-strand cleavage by the L1 EN (red arrowhead) at the 5'TTTT/A consensus (red dotted box) allows L1 mRNA (blue line) to anneal to genomic DNA using its poly(A) tail. Reverse transcriptase activity of L1 ORF2 (green oval) synthesizes L1 cDNA (purple line) using L1 mRNA as template and 3' OH from nicked genomic DNA as primer. Second-strand cleavage (blue arrowhead) occurs 7-20 bp downstream from first-strand cleavage site, creating staggered nicks which are later filled in to form TSDs (blue dotted boxes). Attachment of the L1 cDNA and synthesis of the second strand complete the insertion process. TSD sequences for this diagram are from a 637 bp human L1 element located at chr1:65036188-65036824.
(B) Schematic representation of an NCLI event. Following creation of a genomic double-strand break (red thunderbolt), free-floating L1 mRNA (blue line) attaches to newly separated ends using small stretches of complementary bases. Once gap is bridged, it may be filled in by DNA synthesis by either the L1 RT, cellular repair polymerases or both. L1 insertion thus created lacks structural features of TPRT-mediated insertion.

(Szak et al. 2002). Integration of the newly synthesized cDNA and completion of second-strand synthesis are the remaining steps in the TPRT model; however, the order in which they occur and their exact mechanism remain unclear (Zingler et al. 2005). Apart from the presence of TSDs,

other structural hallmarks of TPRT-mediated L1 insertion include frequent 5' truncations (or truncation/inversions) and intact 3' ends with variable-length A-rich tails (Szak et al. 2002).

In recent years, increasing evidence from cell culture retrotransposition assays suggests that in addition to TPRT-mediated insertion, a second, less-characterized L1 integration pathway may exist that is independent of L1-encoded endonuclease (Cost et al. 2002; Morrish et al. 2002). However, with a few isolated exceptions (Audrezet et al. 2004; Mager et al. 1985; Van de Water et al. 1998), the majority of endonuclease-independent (EN_i) L1 insertions have been recovered in cell lines lacking one or more components of the cellular non-homologous end joining (NHEJ) mechanism, a principal form of DNA double-strand break (DSB) repair (Burma et al. 2006). Consequently, whether EN_i L1 insertion occurs at detectable frequencies when normal DNA repair pathways are functional has been the subject of continued debate (Eickbush 2002; Farkash et al. 2006; Farkash and Prak 2006; Moran and Gilbert 2002; Morrish et al. 2002). Additionally, existing analyses of human genomic L1 elements (Martin et al. 2005; Szak et al. 2002), by focusing solely on TPRT-mediated insertions, have left this question unanswered in a systematic fashion.

In this study, we have utilized computational analyses of the draft sequence of the human genome to recover L1 elements that utilized this alternative pathway of integration (which we term non-classical L1 insertion or NCLI). We report twenty-one loci where L1 elements appear to have inserted without any hallmarks of endonuclease activity. In each case, we verified the

ancestral (i.e., no L1 insertion) state of the loci by resequencing the orthologous positions in the common chimpanzee and rhesus macaque genomes. Overall, our results suggest that NCLI has been active in recent human evolution, and that it provides an alternative "non-selfish" pathway for L1 integration in the human genome. Interestingly, we find that NCLI loci are clustered in gene-rich regions of the genome, in contrast to the distribution of the more common TPRT-mediated L1 insertions. Based on the unique structural features of NCLI-mediated L1 elements, we suggest that this process may be capable of repairing genomic lesions and that it may confer a slight selective advantage to what may be the otherwise deleterious nature of the L1 family. We conclude that non-LTR retrotransposons may have a previously unrecognized role in maintaining human genomic integrity.

Materials and Methods

Computational Screening for Putative EN_i L1 Insertions

To identify NCLI loci in the publicly available human genome, we first downloaded the file chromOut.zip from the UCSC Genome Bioinformatics web site (http://hgdownload.cse.ucsc.edu/downloads.html#human). This archive contains output files from the RepeatMasker (RM) software package (http://www.repeatmasker.org/) run at the –s (sensitive) setting on individual human chromosomes. For this project, the archived files corresponded to RM output from the May 2004 freeze of the human genome (hg17). Next, using our own script, we extracted all L1 insertions from each chromosome. To find elements missing the segment of the 3' UTR normally used during TPRT-mediated insertion, we developed a set of computer programs that scanned the comprehensive list of L1 elements to find all elements truncated beyond 20 bases from the 3' end. We chose the 20 bp truncation limit for two specific reasons. Firstly, from aligning six previously published consensus sequences of relatively young

L1 elements, we found the shortest length of the poly(A) tail to be 13 bp. Secondly, we added a 7 bp window to the 13 bp poly(A) tail to account for the possibility of small internal deletions near the 3' end of the L1 insertions that would mimic the appearance of a 3'-truncated insertion. As RM assigns a size of 6155 bp to full-length L1 elements from subfamilies L1Hs and L1PA2, our initial output files thus contained sets of L1 insertions ending at position 6135 or lower. To verify the effectiveness of this strategy, for each chromosome, we manually inspected sets of 50 loci on either side of this truncation limit. The sets of L1 elements with 3' truncations less than 20 bp did not return any loci matching all of these three criteria; absence of TSDs of any length, absence of a poly(A) tail and significant deviation from the consensus L1 EN cleavage site. Thus, these L1 elements most likely integrated into the genome through traditional TPRT-mediated insertion. As such, after visual inspection of the computational output, all loci that we selected for further experimental verification came from the set of insertions with 3' truncations 20 bp or longer. To further narrow our list to relatively young L1 insertions, we discarded all elements more than 2% diverged from their respective consensus sequences according to the RM algorithm. We rejected all L1 insertions that had TSDs of any length, even if they bordered a 3' truncated element. Our RM output parsing software accounted for L1 elements fragmented by small insertions/deletions and for truncated/inverted L1 insertions, both of which commonly occur during the TPRT process and are sometimes annotated by RM as separate insertions. All the computer programs are available from the authors upon request.

Manual Inspection of Sequence and Verification of Ancestral (Pre-insertion) Status

To confirm the ancestral (i.e. no insertion) stage for computationally recovered NCLI loci, we extracted 10,000 bp of flanking sequence on either side of the L1 element. First, we ran each extracted segment (L1 insertion plus flanking sequence) through RM to verify that the

potential NCLI candidates were not fragments of 3' intact L1 elements separated by large blocks of intervening non-L1 sequence. We then used the BLAT software package (http://www.genome.ucsc.edu/cgi-bin/hgBlat) to construct triple alignments of the human, chimpanzee and rhesus macaque genomes at each locus. Next, we manually inspected each alignment to verify that the 5' and 3' ends of each putative human NCLI event corresponded to either gaps or extra, non-L1 sequence in the ancestral sequence (the presence of non-L1 sequence indicated a deletion in the ancestral genome whose boundaries exactly matched the human L1 insertion). In addition, to further confirm the endonuclease-independent nature of putative NCLI loci, we analyzed them for divergence from the TTTT/A L1-EN cleavage site consensus, based on an earlier analysis of EN site preferences (Morrish et al. 2002). This left us with a final data set of 21 potential NCLI loci that fit all four of the following criteria: 3' truncation, absence of TSDs, absence of a poly(A) tail and significant divergence from the L1-EN consensus.

PCR Amplification and DNA Sequence Analysis of NCLI Loci

To experimentally confirm that these twenty-one loci represented truncated L1 insertions rather than deletions of the 3' UTR, we designed oligonucleotide primers in the non-repetitive sequence flanking the L1 elements and amplified them by PCR on a panel of five primate species (Fig. 4.2), including *Homo sapiens* (HeLa; cell line ATCC CCL-2), *Pan troglodytes* (common chimpanzee; cell line AG06939B), *Gorilla gorilla* (Western lowland gorilla; cell line AG05251), *Macaca mulatta* (Rhesus macaque; cell line NG07098) and *Chlorocebus aethiops* (Green monkey; cell line ATCC CCL70). PCR amplification of NCLI loci was performed in 25 μl reactions using 10-50 ng genomic DNA, 200 nM of each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM MgCl2, 10 mM Tris-HCl (pH 8.4) and 2.5 units

Taq DNA polymerase. The conditions for the PCR were an initial denaturation step of 94° C for four minutes, followed by 32 cycles of one minute of denaturation at 94° C, one minute of annealing at optimal annealing temperature and one minute of extension at 72° C, followed by a final extension step at 72° C for ten minutes. For loci with large insertions or deletions (>2 Kb),

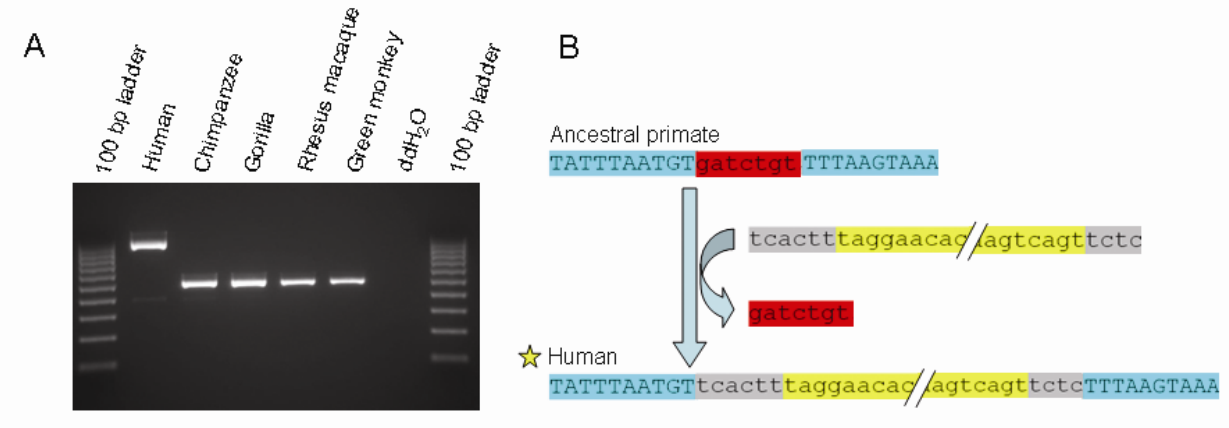


Figure 4.2: Analysis of NCLI elements.
(A) Gel chromatograph of PCR products from a phylogenetic analysis of a human genome-specific NCLI locus (NCLI34). DNA template used in each lane is shown at top.
(B) Schematic diagram of NCLI locus (NCLI53) showing L1 insertion (yellow box) associated with 7 bp deletion of target DNA (red box). Matching flanking sequence is shown as light blue boxes. Grey boxes indicate small segments of non-L1 "filler" DNA at either end of the L1 insertion.

we used *Ex Taq*™ polymerase (TaKaRa) and carried out PCR in 50 µl reactions following the manufacturer's suggested protocol. PCR amplicons were separated on 1% agarose gels, stained with ethidium bromide and visualized using UV fluorescence. Detailed information for each locus including primer sequences, annealing temperature and PCR product sizes is available from the "Publications" section of the Batzer laboratory website (http://batzerlab.lsu.edu). Repetitive DNA may correspond to sites of genome assembly errors, therefore we resequenced all loci from the chimpanzee and rhesus macaque genomes to confirm that the computationally recovered pre-insertion sequence was accurate. Individual PCR products were purified from gels using Wizard® gel purification kits (Promega). Amplicons smaller than 1.3 Kb were cloned into

vectors using TOPO-TA Cloning® kits (Invitrogen) and three colonies were randomly selected and sequenced in both directions using M13 forward and reverse primers to verify that the PCR product matched the computationally recovered sequence. For PCR products larger than 1.3 Kb, gel-purified PCR products were sequenced directly using the respective primers to verify that sequence boundaries matched the computational predictions. All sequencing was performed by the chain termination method (Sanger et al. 1977) on an Applied Biosystems ABI3130*XL* automated DNA sequencer. Analysis of all of the resequenced loci showed that the sequences were exact matches to those in the draft genome sequence assemblies.

Results

A Whole-Genome Scan for Non-Classical L1 Insertions

To analyze the human genome sequence for potential NCLI loci, we combined computational and experimental approaches. First, using RM, we computationally extracted young L1 insertions lacking structural hallmarks of TPRT-mediated "classical" retrotransposition (see Materials and Methods). Next, we constructed triple alignments of the human, chimpanzee and rhesus macaque genomes at these loci to reconstruct the ancestral (i.e., pre-L1 insertion) state and manually inspected the structure of each locus to detect signs of non-TPRT mediated insertion. Finally, we used PCR and resequencing to experimentally verify the sequence architecture for both the post-insertion and pre-insertion states of the loci (Fig. 4.2). Specifically, the loci included in this analysis after experimental confirmation of the computational output possessed all four of the following characteristics: 3' truncation beyond 25 bp (i.e., 5 bp more than the minimum truncation level set during computational screening) relative to the L1HS_3end consensus from the RepeatMaskerLib.embl repetitive element library, downloadable from http://www.girinst.org/repbase/index.html)(Jurka 2000), absence of TSDs of

any length, absence of a poly(A) tail and significant deviation from the consensus L1 EN cleavage site. Structural features of the NCLI loci we extracted using this approach closely mimic EN_i L1 insertions reported in earlier cell-culture analyses (Cost et al. 2002; Farkash et al. 2006; Morrish et al. 2002), further consolidating our hypothesis that they represent products of a similar insertion mechanism in the human genome. We found a total of 21 NCLI loci in the May 2004 freeze of the human genome (hg17)(Table 4.1), of which we were able to recover the pre-insertion site of seven loci from the chimpanzee genome assembly (panTro2; March 2006 freeze) (CSAC 2005) and fourteen loci from the rhesus macaque genome assembly (rheMac2; January 2006 freeze) (RMGSAC 2007). As we were only interested in NCLI loci for which we could verify the pre-insertion sequence, we discarded all L1 insertions that were shared between these three genomes and thus represented older ancestral L1 elements. The L1 elements at NCLI loci ranged between 34 - 4410 bp in length, with a total of 12,018 bp L1 DNA (along with 1365 bp of non-L1 sequence) being captured between the matching 5' and 3' ends of the pre-insertion and post-insertion states. In addition, 18 of 21 NCLI loci were associated with deletions of target site DNA, ranging between 5 bp – 14, 534 bp and totaling 31, 009 bp.

Our estimate of the total number of NCLI events is probably conservative, given that the RM algorithm we used to detect L1 elements, even at its –s (sensitive) setting, is unable to detect insertions smaller than 30 bp. Given that previous cell culture analyses of DSB repair by L1-mediated gene conversion have detected insertion tracts as small as 13 bp (Tremblay et al. 2000), it is quite possible that the number of recent human NCLI events is actually higher than our estimate. Further support for the existence of such "hyphen elements" (Audrezet et al. 2004) in the genome comes from ongoing studies in our lab (Sen, S. K. et al, unpublished data), where we find that TPRT can produce severely 5' truncated L1 and *Alu* insertions with a similar minimum

Table 4.1: Human NCLI loci and insertion site characteristics
In the column for "Lineage", H indicates a NCLI event specific to the human genome, while HC indicates an NCLI event shared between the human and chimpanzee genomes but absent from the rhesus macaque genome

Locus	Coordinates	L1 bp_ins	non-L1 bp_ins	bp_del	L1 seq 5' or 3'?	AT% ±200bp	AT% ±20Kb	Lineage	Intragenic?
NCLI1	chr3:196416805-196421321	4410	107	109	3'	59.5	49.04	H	*C3ORF1*
NCLI3	chr4:67544153-67545039	589	298	1574	3'	63	62.41	H	-
NCLI9	chr17:36395952-36396018	67	0	0	both	60.5	60.86	HC	*KRT40*
NCLI11	chr19:15679181-15680403	1223	0	2867	both	67.5	59.16	H	-
NCLI23	chr2:29588579-29590824	2246	0	17	both	65	56.7	HC	*ALK*
NCLI32	chr4:112069027-112069153	122	5	23	3'	60.5	63.86	HC	-
NCLI33	chr4:60239707-60239936	108	122	2485	3'	65.5	67.18	HC	-
NCLI34	chr4:87186203-87186706	483	21	30	5'	70.5	64	H	*MAPK10*
NCLI38	chr5:51963332-51963788	441	16	1692	3'	53.5	61.77	HC	-
NCLI40	chr6:4414637-4415321	600	85	0	none	58	55.68	HC	-
NCLI47	chr9:108094757-108094921	160	5	8	5'	67.5	60.96	HC	-
NCLI48	chr10:60661882-60662013	34	98	5928	5'	67.5	66.94	HC	*PHYHIPL*
NCLI51	chr11:34668952-34669415	464	0	615	both	58	63.31	H	-
NCLI52	chr12:59792048-59792392	336	9	46	3'	73	65.3	HC	-
NCLI53	chr12:14711194-14711264	61	10	7	none	67	60.16	H	*GUCY2C*
NCLI55	chr13:102553958-102554087	48	62	44	none	52	58.98	HC	-
NCLI57	chr13:80218694-80218899	202	4	5	5'	67.5	64.53	HC	-
NCLI60	chr16:35125561-35125651	86	0	0	both	65.5	63.24	H	-
NCLI61	chr17:3071528-3071879	49	303	14534	5'	68	60.46	HC	-
NCLI64	chr22:45486099-45486153	35	0	1010	both	67.5	51.07	HC	*CERK*
NCLI65	chr22:38619900-38620471	254	318	15	none	63.33	60.75	HC	-
	Total (bp)	12018	1365	31009	**Average**	63.33	60.13		

size (~28-30 bp). As such, it is possible that additional NCLI loci beyond the 21 analyzed here remain undetected in the human genome.

Alignment of L1 segments involved in NCLI events with the full-length consensus sequence of a human-specific L1 subfamily (L1Hs) revealed a tendency to cluster in the downstream half of the L1 consensus, with 18 out of 21 NCLI fragments having 5' truncations 3000 bp or more in addition to their 3' truncations (Fig. 4.3; also see supplemental alignment 1, online). Previous analyses show that most TPRT-mediated genomic L1 insertions are severely 5' truncated (Szak et al. 2002), which may reflect low processivity of the L1 RT or alternatively, host suppression of transcription (Gilbert et al. 2005). The analogous tendency of L1 fragments at NCLI loci to be confined within the downstream half of the element may either be due to the same reasons, or may be moderated by the dynamics of L1 ribonucleoprotein (RNP) positioning at the sites of DSBs (Cost et al. 2002).

Random genomic deletions that remove the 3' ends of classical TPRT-mediated L1 insertions (including the poly(A) tail and the downstream TSD) could mimic the sequence architecture of NCLI loci (Mager et al. 1985). However, by reconstructing the pre-insertion site of all loci (and verifying that the starting point of the 3' flanking sequence remained unchanged before and after the L1 insertion), we effectively minimized the chances of including such events in our data, as it is unlikely that random deletions would repeatedly and precisely remove only L1 sequence, leaving the downstream sequence untouched. Also, for the 18 NCLI loci that were associated with target site deletions, this would require two independent, random deletion events to have taken place at exactly the same position in two separate primate species, which would have vanishingly small probability. The 3' truncated L1 fragment at locus NCLI 40 was not associated with a deletion of target DNA and was followed by an adenosine-rich stretch, making

it possible that an internal deletion had removed the 3' UTR before the poly(A) tail. However, based on the absence of TSDs, high divergence from the L1-EN consensus and presence of non-L1 DNA at both 5' and 3' ends, we decided to include it in our analysis.

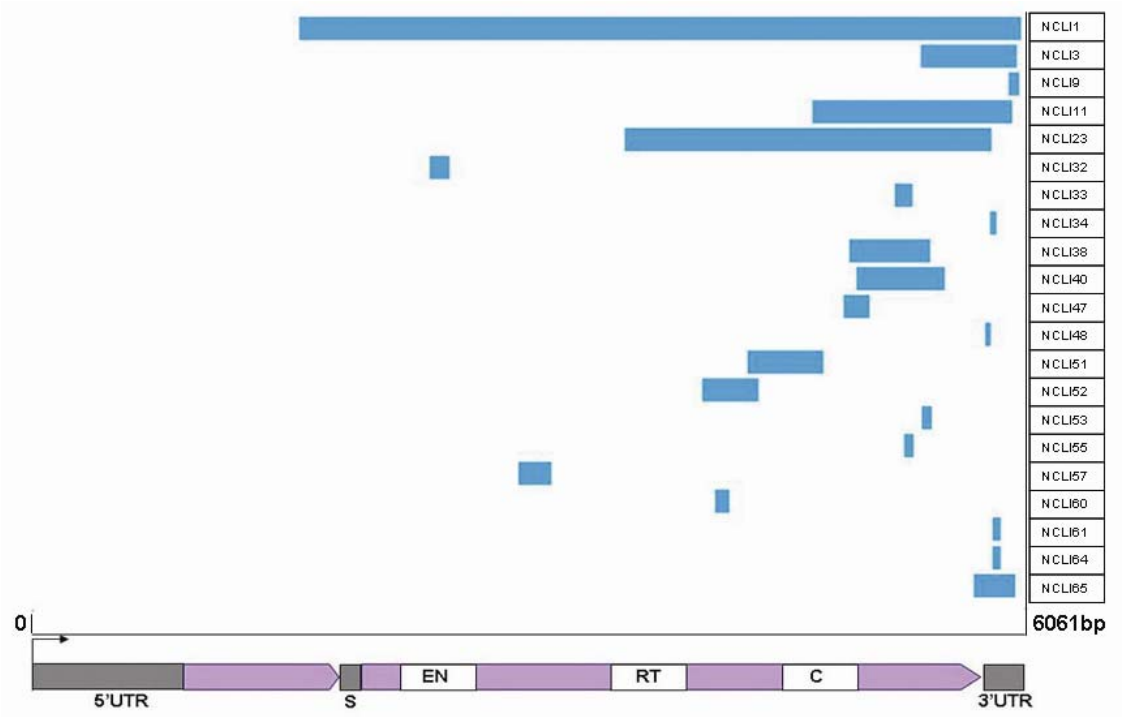


Figure 4.3: Schematic diagram of NCLI L1 element length
Length distribution of L1 segments at the 21 NCLI loci in this analysis along the sequence of a full-length L1 element (L1Hs) as shown by the blue bars.. Location of different domains within the L1 element is shown in the lower panel. Of the non-coding regions (grey boxes) the 5' UTR contains an internal RNA polII promoter, while a 63 bp spacer (S) separates the two ORFs (purple arrows). 40 KDa ORF1 has RNA-binding and nucleic acid chaperone activities, while 150 KDa ORF2 consists of an NH2-terminal endonuclease (EN) domain, a central reverse transcriptase (RT) domain, and a COOH-terminal zinc-knuckle like domain. The extreme 3' end of the 3'UTR consists of a variable poly(A) tail, absent in all 21 NCLI-mediated insertions.

Analysis of Insertion Sites Reveals Divergence from L1-EN Consensus

To find additional evidence supporting our hypothesis that NCLI events were created by an endonuclease-independent mechanism, we inspected all loci for deviations from the 5'-TTTT/A-3' L1-EN consensus cleavage site. Histograms of divergence scores of NCLI events, compared to two other recent analyses of TPRT-mediated L1 insertions (Fig. 4.4), revealed a

marked shift in the maxima towards an increased number of differences. Statistical comparisons of the amounts of deviation from the consensus revealed a highly significant difference between the cleavage site preferences of NCLI loci versus a larger set of 282 recent TPRT-mediated L1 insertions (Morrish et al. 2002) (unpaired t-test assuming unequal variances; $p<0.0001$) (Ruxton 2006), further bolstering our conclusion that breaks in the target DNA at NCLI loci were not products of L1-EN cleavage. Previous *in vitro* analyses have demonstrated that in addition to "preferred" motifs for cleavage, a second set of "atypical" motifs also exists which L1-EN can cleave at lower efficiencies during TPRT (Cost and Boeke 1998; Morrish et al. 2002). However, none of the 21 loci involved insertions into any of these preferred or atypical motifs, further supporting our hypothesis that the NCLI mechanism is independent of L1-EN activity.

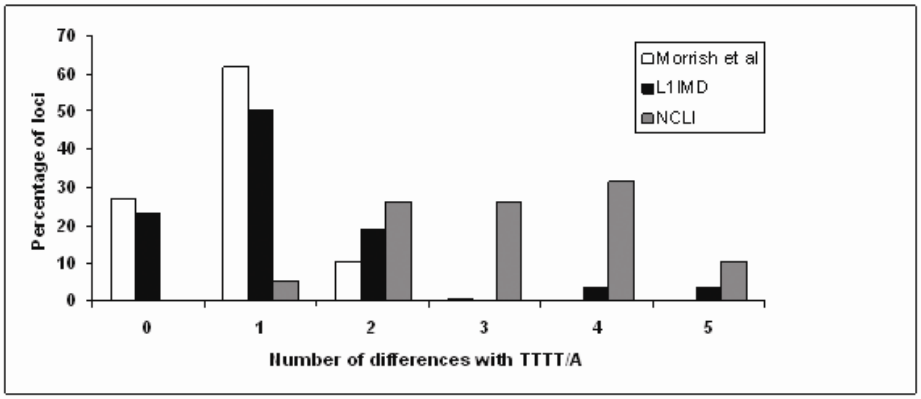


Figure 4.4: L1 cleavage site analysis.
Frequency spectra of deviations from the consensus L1 endonuclease cleavage site (5'-TTTT/A-3'). Two sets of TPRT-mediated insertions are represented along with NCLI events (grey bars): 282 L1_Ta subfamily elements identified in reference 22 (white bars) and 26 human genome-specific L1 insertion-mediated deletions (L1IMDs) identified in reference 10 (black bars).

An analysis of nucleotide composition in the 20 Kb of genomic sequence flanking NCLI loci showed average AT content to be ~ 60% (Table 4.1), which fits well with the global distribution of L1 elements in the human genome (Lander et al. 2001b). Interestingly, even within these AT-rich surroundings, the 200 bp immediately surrounding the breakage sites within the ancestral

genome (100 bp in either direction) showed a further increase in AT content (average of 63.3%). Given that AT-rich DNA is inherently unstable (Chalikian et al. 1999), this trend may reflect the possibility that such stretches in the local sequence architecture, being prone to mechanical or thermodynamic breakage, provide more frequent substrates for NCLI events than nearby GC-rich segments.

Structural Characteristics of NCLI Events

Structurally, NCLI loci closely resemble analogous insertions of non-LTR retrotransposons into preexisting DSBs in cell culture models (Ichiyanagi et al. 2007; Lin and Waldman 2001; Morrish et al. 2002; Teng et al. 1996), which supports our hypothesis that NCLI represents a DSB repair mechanism. Except for occasional 3' transductions which are a byproduct of the TPRT process (Pickeral et al. 2000), classical L1 insertions are rarely associated with insertions of non-L1 DNA. In contrast, 71% of NCLI events (15 out of 21) involved insertions of non-L1 DNA segments of lengths ranging from 4-312 bp along with the L1 DNA (we use the term "extra nucleotides" to denote these segments). Extra nucleotides conjoined to the L1 element were at the 3' and 5' ends at eight and six NCLI loci, respectively, while at three loci such insertions flanked both sides of the L1 element. Closer examination of the extra nucleotides revealed some interesting clues about the possible mechanisms associated with NCLI events, which we discuss below.

At two loci (NCLI1 and NCLI40), fragments of other cellular RNAs appeared to have been co-opted along with the L1 RNA during reverse transcription by the L1 RT. While chimeric L1-U6 snRNA insertions similar to NCLI1 have been previously described (Buzdin et al. 2003; Buzdin et al. 2002), an 18 bp fragment of *GPD2* mRNA was present at the 5' end of NCLI40, providing new evidence that the L1 RT can switch templates between L1 RNA and other cellular

RNAs during the retrotransposition process. At one locus (NCLI3), an intact Alu*Y* element was present at the 5' flank of the L1 insertion. While the Alu*Y* element may have been a later, TPRT-mediated insertion, the absence of TSDs and high divergence from the L1 EN consensus cleavage site suggest that this locus may also represent capture of a nearby *Alu* mRNA during NCLI or an instance of *in vivo* L1-*Alu* RNA recombination.

At two loci, BLAST searches using the extra nucleotides showed evidence for NCLI-mediated inter-chromosomal translocations. At NCLI65 (located on chr. 22), 267 of the 312 extra nucleotides at the 5' flank of the L1 shared significant similarity with a 266 bp stretch on chr. 8 (94% identity; $E = 2e^{-95}$). At the second locus (NCLI40, located on chr. 6), 24 out of 66 extra nucleotides at the 3' end had a near-perfect match on chr. 2 (95% identity; E = 0.059). At a third locus (NCLI34), 11 of 21 extra nucleotides perfectly matched a segment of the *AluJ* consensus sequence. As this *Alu* subfamily has long been inactive in terms of retrotransposition, this may represent the use of an ancient insertion located elsewhere for SDSA-mediated DSB repair (Formosa and Alberts 1986; Nassif et al. 1994); alternatively, the homology could be purely due to chance. At locus NCLI48 (which was associated with a 5928 bp deletion in the ancestral genome), we found additional evidence for the SDSA repair pathway being a component of NCLI. Here, 98 bp extra nucleotide sequence at the 5' end of the human L1 insertion had a highly significant match (96% identity; $E = 4e^{-39}$) to a segment of equal length within the ancestral deletion referred to above. A viable mechanism explaining this structure involves local melting of the double helix within the segment deleted during the NCLI event to provide a transient single-stranded template for repair of the genomic lesion, conforming to the SDSA models described in the earlier studies referred to above. Extra nucleotide stretches at twelve of the 42 junctions (i.e., at either side of the 21 L1 fragments) either did not have

statistically significant BLAST matches in the human genome, or were too small (<15 bp) to draw any definite conclusions. Two junctions (5' end of NCLI33 and 3' end of NCLI61) contained 122 and 303 bp insertions of AT-rich simple repeats, respectively, suggesting that the NCLI process may also contribute to the creation of new microsatellite loci in the human genome, in a manner similar to TPRT-mediated L1 insertion (Ovchinnikov et al. 2001).

In contrast to previous computational analyses that estimate 19-25% of TPRT-mediated L1 insertions in the human genome to be 5'-truncated/inverted (Ostertag and Kazazian 2001b; Szak et al. 2002), only two of the 21 NCLI loci in our analysis showed internal rearrangements within the L1 segment. Interestingly, previous analyses of endonuclease-independent L1 insertions have not recovered any truncated/inverted structures as well (Morrish et al. 2002). In view of these results, we suggest that linearly structured segments in the free-floating L1 mRNA are preferentially captured at the sites of DSBs. Strong support for this hypothesis comes from a previous analysis of ΦK174 DNA fragments transfected into enzymatically created DSBs in a thymidine kinase-deficient mouse cell line, where linear fragments were captured more 9X efficiently than supercoiled segments (Lin and Waldman 2001). Of the two NCLI loci that showed evidence for rearrangement within the L1, NCLI38 was a simple truncation/inversion structure most likely formed by twin priming (Ostertag and Kazazian 2001b). Locus NCLI34, where three consecutive L1 fragments formed a complex structure was more difficult to explain. However, the best BLAST match to the 377 bp highly diverged middle segment (98% identity; E = 0.0) was located downstream on the same chromosome. Thus, our model for this locus suggests an initial truncated/inverted NCLI event followed by a subsequent intra-chromosomal gene conversion which inserted the middle segment. Similar internal rearrangements in L1Hs elements have been documented by a previous analysis (Myers et al. 2002).

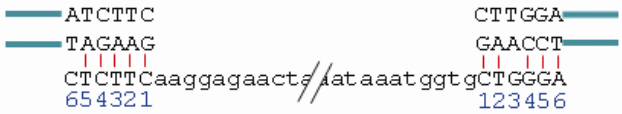
The total amount of deleted sequence between the pre-insertion and post-insertion states of the 21 NCLI events was 31,009 bp, more than twice the 13, 383 bp of combined L1 and non-L1 sequence inserted at the same loci. Of the deleted sequence, almost 50% (14, 534 bp) was associated with a single locus (NCLI61). For this locus, as for all others, we confirmed by both PCR and resequencing that the computationally detected deletion was authentic and matched the draft genome sequence.

Microhomology between Ends of L1 Inserts and Flanking Host DNA

Recent evidence suggests that microhomology between the L1 mRNA and single-stranded overhangs in the genomic DNA flanking the L1-EN cleavage site mediates 5'-end attachment during conventional TPRT, while the 3' end of the mRNA anneals to the nicked DNA through its poly(A) tail (Martin et al. 2005; Zingler et al. 2005). It is possible that a similar mechanism is used for attachment of the L1 RNA to the target DNA during the NCLI process as well. However, to support this assumption for NCLI loci, increased levels of microhomology would have to be present independently at the 5' and 3' ends of the L1 insertion rather than at the 5' end alone. To detect such stretches of higher-than-random complementarity at the ends of a NCLI locus, wherever an exact junction was present between the L1 element and flanking pre-insertion host sequences, we located (a) the 5' and 3' extremities of the L1 insertion with respect to the L1-Hs consensus sequence and; (b) the starting points of 5' and 3'-end flanking sequence (which we identified by aligning the pre-insertion and post-insertion states of the loci) (Fig.4.5A). Next, we isolated 6 bp stretches of sequence extending outwards from these points (i.e., upstream of the 5' end and downstream of the 3' end) in both the L1Hs consensus and flanking sequence and aligned them to count the number of complementary bases (at loci where non-L1 DNA was present at one end of the L1 insertion, we only analyzed the other end). Given

that microhomology-mediated single-strand annealing can resolve DSBs when the extent of complementarity is limited to even one match (Pfeiffer et al. 1994), the high numbers of complementary bases at the L1-genomic DNA junctions (particularly at the first two positions) noticed separately at both the 5' and 3' ends of NCLI loci (Fig. 4.5B) strongly suggest that a similar mechanism is indeed likely to facilitate L1 mRNA binding during NCLI, and further consolidates our hypothesis that NCLI acts as a DSB repair mechanism.

A



B

Locus	5' microhomology						3' microhomology					
	1	2	3	4	5	6	1	2	3	4	5	6
NCLI1	-	-	-	-	-	-	T	T	A	C	A	G
NCLI3	-	-	-	-	-	-	C	T	G	T	T	T
NCLI9	A	A	C	G	A	A	A	A	A	T	A	A
NCLI11	C	T	T	C	C	T	C	A	T	A	A	A
NCLI23	A	T	C	T	G	G	A	T	T	T	C	C
NCLI32	-	-	-	-	-	-	A	A	G	G	A	A
NCLI33	-	-	-	-	-	-	A	A	T	A	T	T
NCLI34*	G	A	G	T	T	A	-	-	-	-	-	-
NCLI38	-	-	-	-	-	-	A	A	G	C	T	C
NCLI47	A	T	C	T	T	T	-	-	-	-	-	-
NCLI48*	A	C	A	A	A	A	-	-	-	-	-	-
NCLI51	C	T	T	C	T	A	C	T	T	G	G	A
NCLI52	-	-	-	-	-	-	G	C	A	T	T	A
NCLI57	T	A	A	A	T	A	-	-	-	-	-	-
NCLI60	A	A	T	A	C	A	A	A	C	T	A	C
NCLI61	T	T	C	A	C	A	-	-	-	-	-	-
NCLI64	C	A	C	T	C	T	G	A	C	T	T	T
Frequency	9/11	7/11	6/11	6/11	5/11	3/11	12/12	11/12	8/12	3/12	5/12	3/12

5' microhomology			3' microhomology		
Position	r	*P*-value	Position	r	*P*-value
1	9	0.000118	1	12	0.000000
2	7	0.006373	2	11	0.000002
3	6	0.026766	3	8	0.002390
4	6	0.026766	4	3	0.258104
5	5	0.080299	5	5	0.103241
6	3	0.258104	6	3	0.258104

Figure 4.5: NCLI microhomology analysis
(A) Complementarity at the 5' and 3' ends of NCLI loci. Note that nucleotide positions are counted in opposite directions at the 5' and 3' ends, based on the first two nucleotides that would anneal during mRNA attachment.
(B) Number of matches at each position and the corresponding *P*-values, that indicate the likelihood of obtaining the observed numbers of matches by chance alone. Bases are highlighted grey if they are complementary to the corresponding nucleotide on the L1 RNA. P-values were calculated based on a binomial probability distribution, where the chance of success (i.e complementary pairing) at each position was ¼ and the chance of failure was ¾.

Genomic Environment of NCLI Events

To characterize the genomic context in which NCLI events occur, we scanned 2 Mb sequence upstream and downstream of each locus for the presence of known or predicted human genes using NCBI MapViewer (http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html). Surprisingly, compared to the vast majority of L1s in the human genome which are located in gene-poor regions (Lander et al. 2001b; Szak et al. 2002), NCLI events were concentrated in areas of relatively high gene density (one gene/83 Kb), compared to both the global gene density in the human genome (one gene/150 Kb)(IHGSC 2004) and the average gene density in the vicinity of TPRT-mediated L1 insertions associated with human genomic deletions (one gene/200 Kb) (Han et al. 2005). In addition, 33% of NCLI loci (seven out of 21) were situated within the introns of known genes (Table 4.1), twice the figure of 13-17% for TPRT-mediated L1 insertions (Szak et al. 2002). Interestingly, when we analyzed the genomic sequences corresponding to the 19 NCLI-mediated deletions in the ancestral (i.e., chimpanzee or rhesus) genomes, we found that at one locus (NCLI61), a model rhesus gene (*LOC721417*) from the olfactory receptor family had been deleted during the L1 insertion process. Although the olfactory receptor gene family is one of the largest in primate genomes with ~1000 members (Young and Trask 2002) and the deletion of a single gene is unlikely to create a significant difference in phenotype, this event further underscores the tendency of NCLI loci to be concentrated in gene-rich areas of the genome.

Discussion

An Alternative Pathway for Non-LTR Retrotransposition in the Human Genome

In this analysis, we address one of the remaining questions in L1 element biology: does an alternative pathway exist for L1 retrotransposition in the human genome? (Moran and Gilbert

2002; Ostertag and Kazazian 2001a) All through the late 1980s until the introduction in 1993 of the TPRT model for insertion of the R2Bm non-LTR retrotransposon in *Bombyx mori* (Luan et al. 1993), it was thought that L1 propagation occurred mainly through fortuitous insertion into DSBs (Edgell et al. 1987; Voliva et al. 1984). Subsequent research established beyond reasonable doubt that the L1 elements use a TPRT-like process as their predominant insertional mechanism (Cost and Boeke 1998; Cost et al. 2002; Feng et al. 1996) and the focus of L1 element biology has since shifted to resolving the unanswered questions of the TPRT model (Babushok et al. 2006; Zingler et al. 2005). Interestingly though, the hypothesis that an alternative, EN_i mechanism acts concurrently with TPRT in the human genome, though often speculated upon (Cost et al. 2002; Morrish et al. 2002), has never been fully investigated. Existing whole-genome analyses of L1 activity have focused solely on TPRT-mediated insertions, and while EN_i L1 retrotransposition has been earlier been detected in cell lines deficient for DNA repair proteins (Morrish et al. 2002), the authors of these studies suggest that, *in vivo* (i.e., when cellular DNA repair mechanisms function normally), such insertions may not be present at detectable frequencies. Thus, the NCLI loci detected in our study represent the first whole-genome analysis of EN_i L1 insertions in a phenotypically normal genetic background that is also subject to selection (i.e., an extant genome). Additionally, we find that the structures of NCLI events recovered *in vivo* closely mirror those previously found *in vitro,* reaffirming the validity of cell culture retrotransposition assays as surrogate models for analyzing retrotransposon biology and determining the impact that these elements have on the genome.

While it remains possible that further NCLI exist in the human genome that cannot be detected using our computational strategy, TPRT-mediated insertions will regardless be several orders of magnitude more frequent. This disparity in scale can be explained by the fundamentally

different natures of the TPRT and NCLI mechanisms. From the retrotransposon point of view, TPRT is an "independent" process, as L1 elements encode both the endonuclease and reverse transcriptase activities required for self-propagation through this mechanism. As such, TPRT-mediated insertion does not have to depend on pre-existing DSBs to provide integration sites. However, in contrast to the independent and organized nature of TPRT, structural features of NCLI loci suggest that it is a more random process, depending entirely on presence of pre-existing DSBs to provide integration sites. Additionally, while only ~2% of human-specific TPRT-mediated L1 insertions create deletions of target genomic DNA (Han et al. 2005), the fact that 86% of NCLI loci (18 out of 21) are associated with genomic deletions would render it a rather inefficient mechanism, had L1 insertion been its sole function. Thus, it is possible that both these processes have co-existed over long periods of time, and while TPRT has doubtless been the primary mode of insertion, certain beneficial features of NCLI have probably contributed to its persistence despite the relative paucity of these events (see below). The observation that at least seven NCLI events are restricted to the human lineage and absent from the chimpanzee and rhesus macaque genomes suggests that this process has been active in recent human genome evolution subsequent to the divergence of human and non-human primates.

Mechanistic Aspects of NCLI Suggest a Role in DNA Repair

The NCLI loci we analyzed may have been produced by three separate mechanisms: (a) capture of nearby L1 mRNAs at the site of DSBs and subsequent reverse transcription (Fig. 4.1B); (b) SDSA-mediated DSB repair in which the free-floating ends of a DSB transiently invade locally melted regions of neighboring double-stranded DNA to provide templates for transcription (Formosa and Alberts 1986; Nassif et al. 1994) and; (c) conventional double-strand break-induced recombination (DSBR) (Liang et al. 1998). Since only three out of 21 NCLI loci

(NCLI11, NCLI23 and NCLI51) involve insertions into pre-existing L1 elements, it is unlikely that conventional DSBR is a mechanism for NCLI, since the presence of sequence homology between the recombining strands is a prerequisite for this model. Of the other two pathways, while theoretically possible, we believe that SDSA is not a preferred mechanism for NCLI. Firstly, the SDSA pathway is highly efficient at minimizing loss of genomic DNA during the patching of DSBs (McVey et al. 2004), which contrasts with the ~31 Kb of genomic deletion detected at the NCLI loci in our analysis. Secondly, although L1 family insertions comprise ~17% of the genome (Lander et al. 2001b), subfamilies which have been active in recent human genome evolution comprise only a small fraction of this figure, while the vast majority of insertions belong to older, extinct subfamilies and have accumulated large numbers of mutations relative to the original consensus sequence (Khan et al. 2006; Mathews et al. 2003; Smit et al. 1995). In this scenario, it is unlikely that the much smaller fraction of recent L1 insertions would be preferentially chosen as templates for SDSA-mediated repair at the 21 NCLI loci in our analysis, which invariably involve relatively young L1 elements with few internal mutations (i.e., <2% divergent by the RM algorithm; see Materials and Methods). However, at two loci (NCLI34 and NCLI48), we did find some evidence that SDSA may play a minor accessory role in the NCLI mechanism (see Results).

Consequently, our preferred model for NCLI is that L1 mRNAs occasionally act as genomic Band-Aids® by bridging pre-existing DSBs in the genome. Given that unrepaired DSBs are among the most lethal forms of DNA damage (Burma et al. 2006; Jackson 2002), it is not surprising that mammalian cells have evolved highly efficient repair pathways capable of patching DSBs with almost any DNA molecule available in the vicinity (Lin and Waldman 2001). Indeed, capture of mobile DNA (including DNA transposons and both LTR and non-LTR

retrotransposons) at the site of genomic DNA lesions seems to be a recurring theme in eukaryotic cells (Ichiyanagi et al. 2007; Lin and Waldman 2001; Morrish et al. 2002; Teng et al. 1996; Yu and Gabriel 1999). In addition, the exceptionally high levels of complementary bases at the L1-host DNA junctions at NCLI loci support our hypothesis of microhomology-mediated L1 mRNA capture between preexisting DSB ends. A recent analysis shows that the L1-EN creates many more genomic DSBs than is required for its own retrotransposition (Gasior et al. 2006), raising an interesting question: are some of these newly created breaks promptly filled in by NCLI? While we consider this to be a possibility, the NCLI loci analyzed in our study all show significant deviations from the L1-EN site, making it unlikely that any of them represent such an occurrence. However, NCLI could be considered a genomic "payoff", through which L1 elements partially compensate for the excess of DSBs that they create. Additional studies of other non-autonomous L1-dependent retrotransposons such as *Alu* and SVA elements will provide further insight into the role these elements may play in NCLI.

Previous analyses have shown that cellular DNA repair proteins used in the non-homologous end-joining (NHEJ) pathway mobilize to the sites of DSBs and compete with the DSBR repair machinery where both systems are available (Drouet et al. 2005; Drouet et al. 2006; Rapp and Greulich 2004). Given that both NCLI and NHEJ are error-prone repair pathways associated with loss of genomic sequence, we consider it quite probable that the NHEJ machinery is co-opted at NCLI loci. Significantly, NHEJ proteins have previously been shown to co-fractionate with non-LTR retrotransposon cDNA intermediates, further supporting the hypothesis that they are involved in the genomic integration of mobile DNA (Downs and Jackson 1999; Downs and Jackson 2004).

Conclusion

In this study, we have demonstrated that NCLI has provided an alternative, endonuclease-independent pathway for L1 integration during human genome evolution, and highlighted its structural differences as compared to the more common and well-characterized TPRT-mediated mode of L1 insertion in the human genome. Based on the sequence architecture of NCLI loci, we propose that this mechanism has been a fortuitous mode for repair of genomic lesions. The distinct nature of the TPRT and NCLI processes suggests that they may have different genomic implications. TPRT-mediated L1 insertions in the human genome, apart from creating large numbers of double-strand breaks, are associated with disruption of functional genes and may be prone to post-insertion ectopic recombination. On the contrary, both the genomic NCLI loci we have detected and similar insertions in previous cell-culture analyses show definite signs of being variants of DSB repair, and seven of the loci we have detected are located within protein-coding genes, breakage within which would otherwise have had direct consequences on the phenotype. Thus, it is interesting to speculate that this "non-selfish" role of NCLI-mediated insertions in maintaining genomic integrity may result in a qualitative difference in the selective regimes acting on the TPRT and NCLI processes (Boissinot et al. 2001). Seven of the NCLI events we have recovered are specific to the human lineage. Assuming the total number of human lineage-specific L1 insertions to be ~1300-1800 (Han et al. 2005; Lee et al. 2007), NCLI thus occurs at the relatively low frequency of 0.5% in the human genome. However, extrapolating these numbers to the larger timescale of the primate radiation, the ~520,000 L1 elements in primate genomes may thus include ~2000 -2800 NCLI events, making this process a significant factor in shaping the architecture of the genome. In our opinion, the finding that both L1 and *Alu* elements in the human genome are capable of acting as *in vivo* molecular Band-

Aids® is significant, as it opens the possibility that active non-LTR retrotransposon families in primate genomes may have a role in maintaining genomic integrity that awaits further characterization.

References

Audrezet, M.P., J.M. Chen, O. Raguenes, N. Chuzhanova, K. Giteau, C. Le Marechal, I. Quere, D.N. Cooper, and C. Ferec. 2004. Genomic rearrangements in the CFTR gene: extensive allelic heterogeneity and diverse mutational mechanisms. *Hum Mutat* **23:** 343-357.

Babcock, M., A. Pavlicek, E. Spiteri, C.D. Kashork, I. Ioshikhes, L.G. Shaffer, J. Jurka, and B.E. Morrow. 2003. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res* **13:** 2519-2532.

Babushok, D.V. and H.H. Kazazian, Jr. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat* **28:** 527-539.

Babushok, D.V., K. Ohshima, E.M. Ostertag, X. Chen, Y. Wang, P.K. Mandal, N. Okada, C.S. Abrams, and H.H. Kazazian, Jr. 2007. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Res* **17:** 1129-1138.

Babushok, D.V., E.M. Ostertag, C.E. Courtney, J.M. Choi, and H.H. Kazazian, Jr. 2006. L1 integration in a transgenic mouse model. *Genome Res* **16:** 240-250.

Bailey, J.A., G. Liu, and E.E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Boissinot, S., P. Chevret, and A.V. Furano. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17:** 915-928.

Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18:** 926-935.

Britten, R.J. and D.E. Kohne. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161:** 529-540.

Brookfield, J.F. 2001. Selection on Alu sequences? *Curr Biol* **11:** R900-901.

Brookfield, J.F. 2005. The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet* **6:** 128-136.

Brosius, J. and S.J. Gould. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci U S A* **89:** 10706-10710.

Burma, S., B.P. Chen, and D.J. Chen. 2006. Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA Repair (Amst)* **5:** 1042-1048.

Burwinkel, B. and M.W. Kilimann. 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277:** 513-517.

Buzdin, A., S. Ustyugova, E. Gogvadze, Y. Lebedev, G. Hunsmann, and E. Sverdlov. 2003. Genome-wide targeted search for human specific and polymorphic L1 integrations. *Hum Genet* **112:** 527-533.

Buzdin, A., S. Ustyugova, E. Gogvadze, T. Vinogradova, Y. Lebedev, and E. Sverdlov. 2002. A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. *Genomics* **80:** 402-406.

Callinan, P.A., J. Wang, S.W. Herke, R.K. Garber, P. Liang, and M.A. Batzer. 2005. Alu Retrotransposition-mediated Deletion. *J Mol Biol* **348:** 791-800.

Campbell, A. 2002. Eubacterial Genomes. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 1024-1039. ASM Press, Washington, D.C.

Carter, A.B., A.H. Salem, D.J. Hedges, C.N. Keegan, B. Kimball, J.A. Walker, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2004. Genome-wide analysis of the human Alu Yb-lineage. *Hum Genomics* **1:** 167-178.

Chalikian, T.V., J. Volker, G.E. Plum, and K.J. Breslauer. 1999. A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques. *Proc Natl Acad Sci U S A* **96:** 7853-7858.

Check, E. 2007. James Watson's genome sequenced. In *Nature News*. Nature Publishing Group.

Chen, F.C. and W.H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444-456.

Clark, A.G., S. Glanowski, R. Nielsen, P.D. Thomas, A. Kejariwal, M.A. Todd, D.M. Tanenbaum, D. Civello, F. Lu, B. Murphy et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302:** 1960-1963.

Conrad, D.F., T.D. Andrews, N.P. Carter, M.E. Hurles, and J.K. Pritchard. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38:** 75-81.

Cordaux, R., J. Lee, L. Dinoso, and M.A. Batzer. 2006a. Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene* **373:** 138-144.

Cordaux, R., J. Lee, L. Dinoso, and M.A. Batzer. 2006b. Recently integrated *Alu* retrotransposons are essentially neutral residents of the human genome. *Gene* **373:** 138-144.

Cost, G.J. and J.D. Boeke. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37:** 18081-18093.

Cost, G.J., Q. Feng, A. Jacquier, and J.D. Boeke. 2002. Human L1 element target-primed reverse transcription in vitro. *Embo J* **21:** 5899-5910.

CSAC. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87.

Dagan, T., R. Sorek, E. Sharon, G. Ast, and D. Graur. 2004. AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res* **32:** D489-492.

Dalton, R. 2006. Sequencers step up to the speed challenge. *Nature* **443:** 258-259.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67:** 183-193.

Deininger, P.L. and M.A. Batzer. 2002. Mammalian retroelements. *Genome Res* **12:** 1455-1465.

Deininger, P.L. and A.M. Roy-Engel. 2002. Mobile Elements in Animal and Plant Genomes. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 1074-1092. ASM Press, Washington, D.C.

Devos, K.M., J.K. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12:** 1075-1079.

Disotell, T.R. and A.J. Tosi. 2007. The monkey's perspective. *Genome Biol* **8:** 226.

Dombroski, B.A., A.F. Scott, and H.H. Kazazian, Jr. 1993. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc Natl Acad Sci U S A* **90:** 6513-6517.

Dorus, S., E.J. Vallender, P.D. Evans, J.R. Anderson, S.L. Gilbert, M. Mahowald, G.J. Wyckoff, C.M. Malcom, and B.T. Lahn. 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell* **119:** 1027-1040.

Downs, J.A. and S.P. Jackson. 1999. Involvement of DNA end-binding protein Ku in Ty element retrotransposition. *Mol Cell Biol* **19:** 6260-6268.

Downs, J.A. and S.P. Jackson. 2004. A means to a DNA end: the many roles of Ku. *Nat Rev Mol Cell Biol* **5:** 367-378.

Drouet, J., C. Delteil, J. Lefrancois, P. Concannon, B. Salles, and P. Calsou. 2005. DNA-dependent protein kinase and XRCC4-DNA ligase IV mobilization in the cell in response to DNA double strand breaks. *J Biol Chem* **280:** 7060-7069.

Drouet, J., P. Frit, C. Delteil, J.P. de Villartay, B. Salles, and P. Calsou. 2006. Interplay between Ku, Artemis, and the DNA-dependent protein kinase catalytic subunit at DNA ends. *J Biol Chem* **281:** 27784-27793.

Ebersberger, I., D. Metzler, C. Schwarz, and S. Paabo. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* **70:** 1490-1497.

Edgell, M.H., S.C. Hardies, D.D. Loeb, W.R. Shehee, R.W. Padgett, F.H. Burton, M.B. Comer, N.C. Casavant, F.D. Funk, and C.A. Hutchison, 3rd. 1987. The L1 family in mice. *Prog Clin Biol Res* **251:** 107-129.

Eickbush, T.H. 2002. Repair by retrotransposition. *Nat Genet* **31:** 126-127.

Enard, W., M. Przeworski, S.E. Fisher, C.S. Lai, V. Wiebe, T. Kitano, A.P. Monaco, and S. Paabo. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418:** 869-872.

Farkash, E.A., G.D. Kao, S.R. Horman, and E.T. Prak. 2006. Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucleic Acids Res* **34:** 1196-1204.

Farkash, E.A. and E.T. Prak. 2006. DNA damage and l1 retrotransposition. *J Biomed Biotechnol* **2006:** 37285.

Feng, Q., J.V. Moran, H.H. Kazazian, Jr., and J.D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905-916.

Formosa, T. and B.M. Alberts. 1986. DNA synthesis dependent on genetic recombination: characterization of a reaction catalyzed by purified bacteriophage T4 proteins. *Cell* **47:** 793-806.

Gasior, S.L., T.P. Wakeman, B. Xu, and P.L. Deininger. 2006. The Human LINE-1 Retrotransposon Creates DNA Double-strand Breaks. *J Mol Biol*.

Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Gilbert, N., S. Lutz, T.A. Morrish, and J.V. Moran. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25:** 7780-7795.

Glowatzki, E. and P.A. Fuchs. 2000. Cholinergic synaptic inhibition of inner hair cells in the neonatal mammalian cochlea. *Science* **288:** 2366-2368.

Gregory, T.R. 2004. Insertion-deletion biases and the evolution of genome size. *Gene* **324:** 15-34.

Hackenberg, M., P. Bernaola-Galvan, P. Carpena, and J.L. Oliver. 2005. The biased distribution of alus in human isochores might be driven by recombination. *J Mol Evol* **60:** 365-377.

Hamaker, H.C. 1978. Approximating the cumulative normal distribution and its inverse. *Appl. Statist.* **27:** 76-77.

Han, K., J. Lee, T.J. Meyer, J. Wang, S.K. Sen, D. Srikanta, P. Liang, and M.A. Batzer. in press. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genetics*.

Han, K., S.K. Sen, J. Wang, P.A. Callinan, J. Lee, R. Cordaux, P. Liang, and M.A. Batzer. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* **33:** 4040-4052.

Hattori, M., A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311-319.

Hattori, Y., N. Okayama, Y. Ohba, Y. Yamashiro, K. Yamamoto, I. Tsukimoto, and M. Kohakura. 1999. The precise breakpoints of a Filipino-type alpha-thalassemia-1 deletion found in two Japanese. *Hemoglobin* **23:** 239-248.

Hayakawa, T., Y. Satta, P. Gagneux, A. Varki, and N. Takahata. 2001. Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene. *Proc Natl Acad Sci U S A* **98:** 11399-11404.

Hedges, D.J. and M.A. Batzer. 2005. From the margins of the genome: mobile elements shape primate evolution. *Bioessays* **27:** 785-794.

Hedges, D.J., P.A. Callinan, R. Cordaux, J. Xing, E. Barnes, and M.A. Batzer. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* **14:** 1068-1075.

Hedges, D.J. and P.L. Deininger. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* **616:** 46-59.

Hinds, D.A., A.P. Kloek, M. Jen, X. Chen, and K.A. Frazer. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38:** 82-85.

Huang, L.S., M.E. Ripps, S.H. Korman, R.J. Deckelbaum, and J.L. Breslow. 1989. Hypobetalipoproteinemia due to an apolipoprotein B gene exon 21 deletion derived by Alu-Alu recombination. *J Biol Chem* **264:** 11394-11400.

Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949-951.

Ichiyanagi, K., R. Nakajima, M. Kajikawa, and N. Okada. 2007. Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. *Genome Res* **17:** 33-41.

IHGSC. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931-945.

Inoue, K. and J.R. Lupski. 2002. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* **3:** 199-242.

Jackson, S.P. 2002. Sensing and repairing DNA double-strand breaks. *Carcinogenesis* **23:** 687-696.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-420.

Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101:** 1268-1272.

Kazazian, H.H., Jr. 2000. L1 retrotransposons shape the mammalian genome. *Science* **289:** 1152-1153.

Kazazian, H.H., Jr. and J.V. Moran. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet* **19:** 19-24.

Khan, H., A. Smit, and S. Boissinot. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* **16:** 78-87.

Kolosha, V.O. and S.L. Martin. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci U S A* **94:** 10155-10160.

Kriegs, J.O., G. Churakov, J. Jurka, J. Brosius, and J. Schmitz. 2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* **23:** 158-161.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001a. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001b. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lee, J., R. Cordaux, K. Han, J. Wang, D.J. Hedges, P. Liang, and M.A. Batzer. 2007. Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390:** 18-27.

Levran, O., N.A. Doggett, and A.D. Auerbach. 1998. Identification of Alu-mediated deletions in the Fanconi anemia gene FAA. *Hum Mutat* **12:** 145-152.

Liang, F., M. Han, P.J. Romanienko, and M. Jasin. 1998. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc Natl Acad Sci U S A* **95:** 5172-5177.

Lin, Y. and A.S. Waldman. 2001. Promiscuous patching of broken chromosomes in mammalian cells with extrachromosomal DNA. *Nucleic Acids Res* **29:** 3975-3981.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.

Liu, W.M., W.M. Chu, P.V. Choudary, and C.W. Schmid. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23:** 1758-1765.

Luan, D.D., M.H. Korman, J.L. Jakubczak, and T.H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72:** 595-605.

Lustig, L.R. and H. Peng. 2002. Chromosome location and characterization of the human nicotinic acetylcholine receptor subunit alpha (alpha) 9 (CHRNA9) gene. *Cytogenet Genome Res* **98:** 154-159.

Mager, D.L., P.S. Henthorn, and O. Smithies. 1985. A Chinese G gamma + (A gamma delta beta)zero thalassemia deletion: comparison to other deletions in the human beta-globin gene cluster and sequence analysis of the breakpoints. *Nucleic Acids Res* **13:** 6559-6575.

Marshall, B., G. Isidro, and M.G. Boavida. 1996. Insertion of a short Alu sequence into the hMSH2 gene following a double cross over next to sequences with chi homology. *Gene* **174:** 175-179.

Martin, S.L. 2006. The ORF1 Protein Encoded by LINE-1: Structure and Function During L1 Retrotransposition. *J Biomed Biotechnol* **2006:** 45621.

Martin, S.L., D. Branciforte, D. Keller, and D.L. Bain. 2003. Trimeric structure for an essential protein in L1 retrotransposition. *Proc Natl Acad Sci U S A* **100:** 13815-13820.

Martin, S.L., W.L. Li, A.V. Furano, and S. Boissinot. 2005. The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenet Genome Res* **110:** 223-228.

Martinez, I., M. Rosa, J.L. Arsuaga, P. Jarabo, R. Quam, C. Lorenzo, A. Gracia, J.M. Carretero, J.M. Bermudez de Castro, and E. Carbonell. 2004. Auditory capacities in Middle Pleistocene humans from the Sierra de Atapuerca in Spain. *Proc Natl Acad Sci U S A* **101:** 9976-9981.

Mathews, L.M., S.Y. Chi, N. Greenberg, I. Ovchinnikov, and G.D. Swergold. 2003. Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am J Hum Genet* **72:** 739-748.

Mathias, S.L., A.F. Scott, H.H. Kazazian, Jr., J.D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808-1810.

Matlik, K., K. Redik, and M. Speek. 2006. L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **2006:** 71753.

McCarroll, S.A., T.N. Hadnott, G.H. Perry, P.C. Sabeti, M.C. Zody, J.C. Barrett, S. Dallaire, S.B. Gabriel, C. Lee, M.J. Daly et al. 2005. Common deletion polymorphisms in the human genome. *Nat Genet* **doi:10.1038/ng1696**.

McCarroll, S.A., T.N. Hadnott, G.H. Perry, P.C. Sabeti, M.C. Zody, J.C. Barrett, S. Dallaire, S.B. Gabriel, C. Lee, M.J. Daly et al. 2006. Common deletion polymorphisms in the human genome. *Nat Genet* **38:** 86-92.

McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36:** 344-355.

McClintock, B. 1956. Intranuclear systems controlling gene action and mutation. *Brookhaven Symp Biol***:** 58-74.

McVey, M., J.R. Larocque, M.D. Adams, and J.J. Sekelsky. 2004. Formation of deletions during double-strand break repair in Drosophila DmBlm mutants occurs after strand invasion. *Proc Natl Acad Sci U S A* **101:** 15694-15699.

Miyamoto, M.M., J.L. Slightom, and M. Goodman. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* **238:** 369-373.

Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530-1534.

Moran, J.V. and N. Gilbert. 2002. Mammalian LINE-1 Retrotransposons and Related Elements. In *Mobile DNA II* (eds. N.L. Craig R. Craigie M. Gellert, and A.M. Lambowitz), pp. 836-869. ASM Press, Washington, D.C.

Moran, J.V., S.E. Holmes, T.P. Naas, R.J. DeBerardinis, J.D. Boeke, and H.H. Kazazian, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87:** 917-927.

Morrish, T.A., N. Gilbert, J.S. Myers, B.J. Vincent, T.D. Stamato, G.E. Taccioli, M.A. Batzer, and J.V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31:** 159-165.

Myerowitz, R. and N.D. Hogikyan. 1987. A deletion involving Alu sequences in the beta-hexosaminidase alpha-chain gene of French Canadians with Tay-Sachs disease. *J Biol Chem* **262:** 15396-15399.

Myers, J.S., B.J. Vincent, H. Udall, W.S. Watkins, T.A. Morrish, G.E. Kilroy, G.D. Swergold, J. Henke, L. Henke, J.V. Moran et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71:** 312-326.

Nassif, N., J. Penney, S. Pal, W.R. Engels, and G.B. Gloor. 1994. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Mol Cell Biol* **14:** 1613-1625.

Newman, T.L., E. Tuzun, V.A. Morrison, K.E. Hayden, M. Ventura, S.D. McGrath, M. Rocchi, and E.E. Eichler. 2005. A genome-wide survey of structural variation between human and chimpanzee. *Genome Res* **15:** 1344-1356.

Ostertag, E.M. and H.H. Kazazian, Jr. 2001a. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35:** 501-538.

Ostertag, E.M. and H.H. Kazazian, Jr. 2001b. Twin priming: a proposed mechanism for the creation of inversions in l1 retrotransposition. *Genome Res* **11:** 2059-2065.

Otieno, A.C., A.B. Carter, D.J. Hedges, J.A. Walker, D.A. Ray, R.K. Garber, B.A. Anders, N. Stoilova, M.E. Laborde, J.D. Fowlkes et al. 2004. Analysis of the Human Alu Ya-lineage. *J Mol Biol* **342:** 109-118.

Ovchinnikov, I., A.B. Troxel, and G.D. Swergold. 2001. Genomic Characterization of Recent Human LINE-1 Insertions: Evidence Supporting Random Insertion. *Genome Res* **11:** 2050-2058.

Petrov, D.A. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* **17:** 23-28.

Pfeiffer, P., S. Thode, J. Hancke, and W. Vielmetter. 1994. Mechanisms of overlap formation in nonhomologous DNA end joining. *Mol Cell Biol* **14:** 888-895.

Pickeral, O.K., W. Makalowski, M.S. Boguski, and J.D. Boeke. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* **10:** 411-415.

Quentin, Y. 1992. Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. *Nucleic Acids Res* **20:** 487-493.

Rapp, A. and K.O. Greulich. 2004. After double-strand break induction by UV-A, homologous recombination and nonhomologous end joining cooperate at the same DSB if both systems are available. *J Cell Sci* **117:** 4935-4945.

Richardson, C. and M. Jasin. 2000. Coupled homologous and nonhomologous repair of a double-strand break preserves genomic integrity in mammalian cells. *Mol Cell Biol* **20:** 9068-9075.

RMGSAC. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222-234.

Rohlfs, E.M., N. Puget, M.L. Graham, B.L. Weber, J.E. Garber, C. Skrzynia, J.L. Halperin, G.M. Lenoir, L.M. Silverman, and S. Mazoyer. 2000. An Alu-mediated 7.1 kb deletion of BRCA1 exons 8 and 9 in breast and ovarian cancer families that results in alternative splicing of exon 10. *Genes Chromosomes Cancer* **28:** 300-307.

Rothberg, P.G., S. Ponnuru, D. Baker, J.F. Bradley, A.I. Freeman, G.W. Cibis, D.J. Harris, and D.P. Heruth. 1997. A deletion polymorphism due to Alu-Alu recombination in intron 2 of the retinoblastoma gene: association with human gliomas. *Mol Carcinog* **19:** 69-73.

Rudiger, N.S., N. Gregersen, and M.C. Kielland-Brandt. 1995. One short well conserved region of Alu-sequences is involved in human gene rearrangements and has homology with prokaryotic chi. *Nucleic Acids Res* **23:** 256-260.

Ruxton, G.D. 2006. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology* **17:** 688-690.

Sakharkar, M.K., V.T. Chow, and P. Kangueane. 2004. Distributions of exons and introns in the human genome. *In Silico Biol* **4:** 387-393.

Salem, A.H., J.S. Myers, A.C. Otieno, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003. LINE-1 preTa elements in the human genome. *J Mol Biol* **326:** 1127-1146.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74:** 5463-5467.

Schmid, C.W. 1998. Does SINE evolution preclude Alu function? *Nucleic Acids Res* **26:** 4541-4550.

Schmid, C.W. 2003. Alu: a parasite's parasite? *Nat Genet* **35:** 15-16.

Sen, S.K., K. Han, J. Wang, J. Lee, H. Wang, P.A. Callinan, M. Dyer, R. Cordaux, P. Liang, and M.A. Batzer. 2006. Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* **79:** 41-53.

Sen, S.K., C.T. Huang, K. Han, and M.A. Batzer. 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35:** 3741-3751.

Shendure, J., R.D. Mitra, C. Varma, and G.M. Church. 2004. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5:** 335-344.

Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* **6:** 743-748.

Smit, A.F., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246:** 401-417.

Sorek, R., G. Ast, and D. Graur. 2002. Alu-containing exons are alternatively spliced. *Genome Res* **12:** 1060-1067.

Sorek, R., G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur, and G. Ast. 2004. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell* **14:** 221-231.

Speek, M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol* **21:** 1973-1985.

Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Szabo, Z., S.A. Levi-Minzi, A.M. Christiano, C. Struminger, M. Stoneking, M.A. Batzer, and C.D. Boyd. 1999. Sequential loss of two neighboring exons of the tropoelastin gene during primate evolution. *J Mol Evol* **49:** 664-671.

Szak, S.T., O.K. Pickeral, W. Makalowski, M.S. Boguski, D. Landsman, and J.D. Boeke. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052.

Teng, S.C., B. Kim, and A. Gabriel. 1996. Retrotransposon reverse-transcriptase-mediated repair of chromosomal breaks. *Nature* **383:** 641-644.

Tremblay, A., M. Jasin, and P. Chartrand. 2000. A double-strand break in a chromosomal LINE element can be repaired by gene conversion with various endogenous LINE elements in mouse cells. *Mol Cell Biol* **20:** 54-60.

Tvrdik, T., S. Marcus, S.M. Hou, S. Falt, P. Noori, N. Podlutskaja, F. Hanefeld, P. Stromme, and B. Lambert. 1998. Molecular characterization of two deletion events involving Alu-sequences, one novel base substitution and two tentative hotspot mutations in the hypoxanthine phosphoribosyltransferase (HPRT) gene in five patients with Lesch-Nyhan syndrome. *Hum Genet* **103:** 311-318.

Van de Water, N., R. Williams, P. Ockelford, and P. Browett. 1998. A 20.7 kb deletion within the factor VIII gene associated with LINE-1 element insertion. *Thromb Haemost* **79:** 938-942.

Varki, A. and T.K. Altheide. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res* **15:** 1746-1758.

Venter, J.C. M.D. Adams E.W. Myers P.W. Li R.J. Mural G.G. Sutton H.O. Smith M. Yandell C.A. Evans R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.

Voliva, C.F., S.L. Martin, C.A. Hutchison, 3rd, and M.H. Edgell. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. *J Mol Biol* **178:** 795-813.

Watanabe, H., A. Fujiyama, M. Hattori, T.D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382-388.

Wei, W., N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian, J.D. Boeke, and J.V. Moran. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21:** 1429-1439.

Wheelan, S.J., Y. Aizawa, J.S. Han, and J.D. Boeke. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15:** 1073-1078.

Wildman, D.E., M. Uddin, G. Liu, L.I. Grossman, and M. Goodman. 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. *Proc Natl Acad Sci U S A* **100:** 7181-7188.

Young, J.M. and B.J. Trask. 2002. The sense of smell: genomics of vertebrate odorant receptors. *Hum Mol Genet* **11:** 1153-1160.

Yu, X. and A. Gabriel. 1999. Patching broken chromosomes with extranuclear cellular DNA. *Mol Cell* **4:** 873-881.

Zimmerly, S., H. Guo, P.S. Perlman, and A.M. Lambowitz. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* **82:** 545-554.

Zingler, N., U. Willhoeft, H.P. Brose, V. Schoder, T. Jahns, K.M. Hanschmann, T.A. Morrish, J. Lower, and G.G. Schumann. 2005. Analysis of 5' junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* **15:** 780-789.

CHAPTER FIVE:

SUMMARY

The predominant role of retrotransposons in human genome plasticity is widely accepted, regardless of whether the elements themselves are construed as genomic "mischief-makers" (Hedges and Deininger 2007) or misunderstood sources of useful genetic variation (Brookfield 2005; Sorek 2007). Certain topics are likely to be hotly debated for some time to come, such as the nature of selection acting on the *Alu* and L1 families (Boissinot et al. 2001; Brookfield 2001; Cordaux et al. 2006). Other questions may be answered in the near future, as the timeframe for personal genome sequencing shrinks from years to months to weeks. As a greater collection of sequenced human genomes forms, the extent of between-individual genetic diversity mediated by mobile elements will become clear (Batzer and Deininger 2002; Wang et al. 2006). Components of this diversity such as "private" *Alu* or L1 insertions and recombination-mediated deletion variants unevenly distributed among human populations may reveal an evolutionary and biomedical significance for these families beyond the existing knowledge base of retrotransposon-mediated diseases recovered from individual patients and tumor samples (Deininger and Batzer 1999). As a preparatory study for these future analyses, in this dissertation, I have examined some aspects of genome remodeling directly linked to retrotransposon activity in the human lineage with the last 5-6 million years.

In chapter two, we located and quantified existing human-specific and chimpanzee-specific L1 insertion-mediated deletions (L1IMDs). We identified lineage-specific L1IMD candidates by computational comparison of the complete human and common chimpanzee genome sequences and confirmed these loci by resequencing. We found that L1 insertions are directly responsible for the removal of ~18 Kb of human genomic sequence and ~15 Kb of chimpanzee genomic sequence within the past 4-6 million years and may have generated over 11,000 deletion events during the primate radiation, removing up to 7.5 Mb of DNA in the

process. We propose a mechanism to explain the correlation of size of the L1 insertion with the size of the deletion it causes and suggest models for the formation of L1 truncation/inversion structures during the deletion process. From an evolutionary perspective, the absence in our data of large L1IMDs found in cell culture analyses (up to 71 kb)(Gilbert et al. 2002; Symer et al. 2002) suggests that such deletions are most likely removed from the population by natural selection.

Chapter three presents a global analysis of *Alu* recombination-mediated deletion (ARMD) in the human genome, a process previously characterized only through isolated somatic loci associated with genetic diseases (Deininger and Batzer 1999). By identifying 492 deletions in the human lineage, more than half of which were located inside genes, we demonstrated that the importance of ARMD extends beyond the causation of genetic disease. The rate of sequence loss by ARMD (~400 kb within the last 4-6 million years), along with preliminary results from an ongoing analysis in our laboratory of L1 recombination-mediated deletion in the human genome, leads us to suggest that non-allelic homologous recombination between *Alu* and L1 elements mediates human genomic deletion on a scale that counteracts (at least partially) unilateral genome size expansion through insertion of new elements (Liu et al. 2003). A recent analysis of ARMD in the chimpanzee genome lends additional support to this hypothesis (Han et al. in press). Interestingly, in both the human and chimpanzee genomes, the majority of ARMD loci seem consistently cluster intragenically, and in a number of cases, human and chimpanzee orthologs of functional genes differ as a result of exons being deleted from one of the genomes through ARMD. As such, it is not implausible to suggest that after the human-chimpanzee divergence from a common ancestor, a portion of lineage-specific changes in phenotype may have been mediated by the ARMD process. Though most ARMD loci persisting over

evolutionary timescales would be expected to be either neutral or only mildly deleterious, it is entirely possible that some of the recombination loci we detected represent instances where the high level of sequence identity maintained between closely spaced Alu elements has been used by the cellular homologous recombination repair (HRR) machinery to patch a DNA double-strand break (DSB) located between the elements.

In chapter four, we find evidence for an endonuclease-independent pathway for L1 integration in the human genome that may have a role in DSB repair. Elements integrating through this non-classical L1 insertion (NCLI) mechanism have distinct structural differences from the more common and well-characterized TPRT-mediated L1 insertions in the human genome. A previous analysis has shown that TPRT-mediated L1 insertions in the human genome are subject to negative selection (Boissinot et al. 2001). However, NCLI, being a random and "non-selfish" insertion pathway, it is possible that elements inserting through this mechanism operate under a different selective regime from TPRT-mediated insertions. Interestingly, ongoing studies in our lab indicate that analogous to NCLI, *Alu* elements in the human genome are also capable of inserting without endonuclease activity and acting as *in vivo* "molecular Band-Aids®" (D. Srikanta et al., unpublished data). In our opinion, both these findings are highly significant, as the relative advantage conferred on a few L1 and *Alu* loci that repair lethal DSBs and ensure cell survival could partially offset the mildly deleterious effect of the larger numbers of TPRT-mediated insertions and could contribute to the persistence of these families in primate genomes.

In conclusion, it is long been evident that retrotransposons are dynamically influencing the structure and function of primate genomes. In this dissertation, I have analyzed some of the structural aspects of retrotransposon-mediated plasticity in the human genome and quantified them in an evolutionary timeframe. As the field of personal genomics gathers momentum, the

biomedical importance of these mechanisms in creating individual genomic diversity will come into view, hopefully in the near future.

References

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18:** 926-935.

Brookfield, J.F. 2001. Selection on Alu sequences? *Curr Biol* **11:** R900-901.

Brookfield, J.F. 2005. The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet* **6:** 128-136.

Cordaux, R., J. Lee, L. Dinoso, and M.A. Batzer. 2006. Recently integrated *Alu* retrotransposons are essentially neutral residents of the human genome. *Gene* **373:** 138-144.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67:** 183-193.

Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Han, K., J. Lee, T.J. Meyer, J. Wang, S.K. Sen, D. Srikanta, P. Liang, and M.A. Batzer. in press. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genetics*.

Hedges, D.J. and P.L. Deininger. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* **616:** 46-59.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.

Sorek, R. 2007. The birth of new exons: Mechanisms and evolutionary consequences. *Rna* **13:** 1603-1608.

Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.

Wang, J., L. Song, D. Grover, S. Azrak, M.A. Batzer, and P. Liang. 2006. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat* **27:** 323-329.

APPENDIX

LETTERS OF PERMISSION

OXFORD
UNIVERSITY PRESS

Rights and New Business Development, Journals
Great Clarendon Street,
Oxford OX2 6DP, UK
Telephone: 44 [0] 1865 354779
Fax: 44 [0] 1865 353485
Email: gemma.puntis@oxfordjournals.org
Ref: FP/NARESE/Sen/GP/0607

11/06/07

Shurjo Kumar Sen
Batzer Lab
Biological Computation & Visualization Center
Louisiana State University
Baton Rouge, LA 70803
USA

Dear Shurjo Kumar Sen,

RE: Nucleic Acids Research, Vol. 33 2005, pp. 4040-52, Han, K. et al

Thank you for your email dated 29 May, requesting permission to reprint the above material. Our permission is granted without fee to reproduce the material, as you are the original author.

Use of the **article** is restricted to republication in your dissertation, available in *print* format only, to be used only in the *English* language. This permission is limited to this particular use and does not allow you to use it elsewhere or in any other format other than specified above.

Please include a credit line in your publication citing full details of the Oxford University Press publication which is the source of the material and by permission of Oxford University Press or the sponsoring society if this is a society journal.

If the credit line or acknowledgement in our publication indicates that material including any illustrations/figures etc was drawn or modified from an earlier source it will be necessary for you to also clear permission with the original publisher. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

Please do not hesitate to contact me if I can be of any further assistance.

Yours sincerely,

G. Puntis

Gemma Puntis
Journals Rights Assistant

University of Chicago Press
Journal Permissions Department
1427 East 60th Street
Chicago, IL 60628
Phone: 773-834-1884

Permission Grant

SHURJO KUMAR SEN
LOUISIANA STATE UNIVERSITY
BATZER LAB
LABORATORY OF COMPARTIVE GENOMICS
BATON ROGUE, LA 70803

Date: May 29, 2007
Grant Number: 101308
Request Date: 5-25-2007
Reference Number: 0049531476

Dear Requester:

Thank you for your request for permission to use material from the publications(s) of the University of Chicago Press. Permission is granted for use as stated below. Unless specifically granted below, this permission does not allow the use of our material in any other edition or by any additional means of reproduction including (by way of example) motion pictures, sound tapes, electronic formats, and phonograph records; nor does this permission cover book clubs, translations, abridgment, or selections which may be made of the publication. No subsequent use may be made without additional approval. This permission is subject to the following terms:

1) On each copy of the selection full credit must be given to the book or journal, to the author (as well as to the series, editor, or translator, if any), and to the University of Chicago as publisher. In addition, the acknowledgement must include the identical copyright notice as it appears in our publication.

2) This permission does not apply to any part of the selection which is independently copyrighted or which bears a separate source notation. The responsibility for determining the source of the material rests with the prospective publisher of the quoted material.

3) This permission grant for the materials listed on the invoice below is provided GRATIS.

4) This Permission covers publication of one edition of the work up to 250 copies.

5) Permission granted is non-exclusive and, unless otherwise stated, is valid THROUGHOUT THE WORLD IN THE ENGLISH LANGUAGE ONLY.

6) Author approval is not required.

7) Permission is granted GRATIS.

Reference	ISBN	Materials	ExtPrice
0049531476	2002	SEN/HAN/WANG,AMERICAN JRNL HUMAN GENETICS PERMS. Pages 41 to 53. "HUMAN GENOMIC DELETIONS MEDIATED BY RECOMBINATION BETWEEN ALU ELEMENTS" 79(1), 2006	$0.00

Order Total:	$0.00
Handling:	$0.00
Tax:	$0.00
Sub Total:	$0.00
Payments:	$0.00
Balance Due:	**$0.00**

For Use In:
TO APPEAR IN: PH.D DISSERTATION BY SHURJO KUMAR SEN

OXFORD
UNIVERSITY PRESS

Rights and New Business Development, Journals
Great Clarendon Street,
Oxford OX2 6DP, UK
Telephone: 44 [0] 1865 354779
Fax: 44 [0] 1865 353485
Email: gemma.puntis@oxfordjournals.org
Ref: FP/NARESE/Sen/GP/0707

12/07/07

Mr S K Sen
107 Life Sciences Building
Louisiana State University
Baton Rouge, LA 70803
USA

Dear Mr Sen,

RE: Nucleic Acids Research, Vol. 35 2007, pp. 3741-51

Thank you for your email dated 2 July, requesting permission to reprint the above material. Our permission is granted without fee to reproduce the material, as you are the original author.

Use of the **article** is restricted to republication in your thesis, available in *print* format only, to be used only in the *English* language. This permission is limited to this particular use and does not allow you to use it elsewhere or in any other format other than specified above.

Please include a credit line in your publication citing full details of the Oxford University Press publication which is the source of the material and by permission of Oxford University Press or the sponsoring society if this is a society journal.

If the credit line or acknowledgement in our publication indicates that material including any illustrations/figures etc was drawn or modified from an earlier source it will be necessary for you to also clear permission with the original publisher. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

Please do not hesitate to contact me if I can be of any further assistance.

Yours sincerely,

Gemma Puntis
Journals Rights Assistant

VITA

Shurjo Kumar Sen is the son of Dr. Sudipta Kumar Sen and Mrs. Nupur Sen. He was born in London in 1979, but moved to India with his parents in 1982, where he lived in the states of Assam and West Bengal. Shurjo graduated with a Bachelor of Science (Honours) degree in zoology from Presidency College, Calcutta, in 2001 and a Master of Science degree in zoology from the University of Calcutta in 2003, after which he worked with Dr. Sumit Homechaudhuri for a year as a Research Assistant at the University of Calcutta. Subsequently, he began his doctoral research in the fall of 2004 in the Department of Biological Sciences at Louisiana State University in Baton Rouge, Louisiana, under the direction of Professor Mark A. Batzer. Mr. Sen will graduate with the degree of Doctor of Philosophy in May, 2008.